A theoretical framework for computer models of cooperative dialogue, acknowledging multi-agent conflict

Thesis

A Theoretical Framework for Computer Models of Cooperative Dialogue, Acknowledging Multi-agent Conflict.

Julia Rose Galliers

Thesis submitted in partial fulfillment of requirements for Ph.D in Artificial Intelligence from the department of Psychology, Open University, September 1988.

ABSTRACT

This thesis describes a theoretical framework for modelling cooperative dialogue. The linguistic theory is a version of speech act theory adopted from Cohen and Levesque, in which dialogue utterances are generated and interpreted pragmatically in the context of a theory of rational interaction. The latter is expressed as explicitly and formally represented principles of rational agenthood and cooperative interaction. The focus is the development of strategic principles of multi-agent interaction as such a basis for cooperative dialogue. In contrast to the majority of existing work, these acknowledge the positive role of conflict to multi-agent cooperation, and make no assumptions regarding the benevolence and sincerity of agents. The result is a framework wherein agents can resolve conflicts by negotiation. It is a preliminary stage to the future building of computer models of cooperative dialogue for both HCI and DAI, which will therefore be more widely and generally applicable than those currently in existence.

The theory of conflict and cooperation is expressed in the different patterns of mental states which characterise multi-agent conflict, cooperation and indifference as three alternative postural relations. Agents can recognise and potentially create these. Dialogue actions are the strategic tools with which mental states can be manipulated, whilst acknowledging that agents are autonomous over their mental states; they have control over what they acquire and reveal in dialogue. Strategic principles of belief and goal adoption are described in terms of the relationships between autonomous agents' beliefs, goals, preferences, and interests, and the relation of these to action. Veracity, mendacity, concealing and revealing are defined as properties of acts. The role of all these elements in reasoning about dialogue action and conflict resolution, is tested in analyses of two example dialogues; a record of a real trade union negotiation and an extract from "Othello" by Shakespeare.

ACKNOWLEDGEMENTS

I would like to thank the following people for their assistance during the time I have spent studying for my Ph.D.:

First and foremost, my supervisor, George Kiss, for hours of discussion, interest and support. I have gained so much from him, whilst also enjoying his company and friendship.

Secondly, colleagues such as Steve Draper, Jerry Hobbs, Peter Jackson, Ian Morley, Steve Pulman, Han Reichgelt, Nigel Shadbolt, and Richard Williams, for their helpful comments on documents, discussions of ideas and correspondance. Also, Dave Wilson for his drawing of the robot Willie.

I would like to thank my children, Ben and Ruth for just being, and my mother, Vera Coppard for ever constant support and encouragement. Particularly, I would like to thank Gillian Carter as a most valued friend and sister, for her support in reading, discussing, listening and encouraging, and also for her literary expertise with respect to "Othello".

Finally, I would like to thank the Economic and Social Research Council for financial support.

Julia Rose Galliers

Aspects of the thesis entitled "A Theoretical Framework for Computer Models of Cooperative Dialogue, Acknowledging Multi-agent Conflict", can be found in the following publications:

1. A Strategic Framework for Multi-Agent Cooperative Dialogue. Proceedings of the Eighth European Conference on Artificial Intelligence, Munich, pp 415-420, August, 1988.

2. A Definition of Cooperation (but not Benevolence), for multi-agent planning. Report of the Alvey-IKBS-Expert Systems Research Theme, 7th Planning SIG, Martlesham, Ipswich, November 1987.

3. A Theoretical Framework for Modelling Conflict and Cooperation in Dialogue. British Psychological Society symposium, London, December 1986.

None of the above comprise any particular section of the thesis. Publication 1 is a recent overview of the entire thesis, publication 2 is an expression of some of the views contained in chapter 5, and publication 3 is an early and general discussion of the issues to which the thesis relates.

TABLE OF CONTENTS

PAGE NO.

INTRODUCTION

Research Aims....1
Motivations....1
Original Contribution....2
Theoretical Background....3
Methodology and Testing....3
Disclaimer....4
Thesis Outline....5
Illustrating the Framework....6

CHAPTER 1: A PRAGMATIC THEORY OF DIALOGUE

1.1 Introduction....10
1.2 Approaches to modelling dialogue in artificial intelligence....11
1.2.1 Aims and applications....11
1.2.2 Methods and Theories....12
1.2.3 Speech as planned action....13
1.3 Speech act theory....15
1.3.1 Theoretical background and historical development....15
1.3.2 Formalising speech act theory for computational models- the early work....20
1.3.3 What was wrong with the early work?....24
1.3.4 Alternative approaches/theories....26
1.3.5 Formalising speech act theory for computational models- recent developments from Cohen and Levesque, and Perrault....29
1.4 Conclusions....33

CHAPTER 2: A THEORY OF RATIONAL AGENTHOOD

2.1 Introduction....34

2.2 Rational, intelligent agenthood - the agent model....35

2.2.1 Representing mental states....36

2.2.2 Knowledge and belief....37

2.2.2.1 Belief revision....40

2.2.3 Preference....42

2.2.4 Wants and goals....44

2.2.5 Interests....47

2.2.6 Intention....48

2.2.7 Rational action....49

CHAPTER 3: A THEORY OF COOPERATIVE, MULTI-AGENT INTERACTION

3.1 Introduction....55

3.2 Rational, intelligent agents in interaction - the nature of multi-agent systems....55

3.2.1 Mutual beliefs and common knowledge....57

3.2.2 Propositional postures - conflict, cooperation and indifference....59

3.2.2.1 Existing notions of cooperation in artificial intelligence....61

3.2.2.2 Criticisms of existing notions of cooperation....65

3.2.3 Dialogue and posture - the nature of negotiation....68

3.2.4 The autonomy of interacting agents - control of information....71

3.2.5 Strategic interaction....76

3.2.5.1 The relevance of game theory....76

3.2.5.2 Strategic rationality....80

3.3 Conclusions....83

CHAPTER 4: A FORMAL APPROACH TO MODELLING THEORIES OF AGENTS

4.1 Introduction....85
4.2 Why a formal approach?....86
4.3 Formal approaches to the modelling of agents....90
4.3.1 Knowledge and belief representation....90
4.3.1.1 Modal logics and possible-worlds semantics - the "classical" model....92
4.3.1.2 Semantic approaches and logical omniscience....93
4.3.2 Moore's model - reasoning about knowledge and action....96
4.3.3 Cohen and Levesque's model - reasoning about intention and commitment....97
4.4 A formal model of rational agenthood....99
4.4.1 Syntax....100
4.4.2 Semantics....101
4.4.3 Properties taken directly from Cohen and Levesque's model....104
4.4.3.1 Properties of acts and temporal modalities....104
4.4.3.2 Properties of attitudes....106
4.4.1 Some additional properties of the model....109
4.4.4.1 Preference....110
4.4.4.2 Interests....112
4.5 Conclusions....114

CHAPTER 5: CONFLICT, COOPERATION AND INDIFFERENCE

5.1 Introduction....116
5.2 Justification of social theories for computer applications....117
5.3 The nature of conflict....118
5.3.1 A formal definition of multi-agent conflict....120
5.4 The nature of cooperation....124
5.5 The nature of indifference....127
5.6 Mixed postures in the multi-agent system....128
5.7 The role of conflict in multi-agent systems....129
5.7.1 The positive functions of conflict....129
5.7.2 Conflict, autonomy, and the evolution of cooperation....131
5.8 Conclusions....134

CHAPTER 6: CONTROL OF INFORMATION, STRATEGIC INTERACTION AND THE AUTONOMY OF AGENTS

6.1 Introduction....136
6.2 Information....137
6.3 Control over information acquired in dialogue....139
6.3.1 Conditions for belief and goal adoption in strategic interaction....140
6.3.1.1 An introductory discussion of the issues - a comparison with Cohen and Levesque's approach....140
6.3.1.2 The strategic approach....146
6.4 Control over information revealed in dialogue....157
6.4.1 Expression - a relation between an utterance and a mental state....158
6.4.1.1 Definitions of veracity, mendacity, revealing and concealing.159
6.4.2 Practical implications....161
6.5 Communicative acts....162
6.5.1 The strategic approach....162
6.5.2 The strategic approach contrasted with Cohen and Levesque's - an illustrative summary....165
6.6 Strategic objectives....166
6.7 Conclusions....167

CHAPTER 7: TESTING AND EVALUATING THE FRAMEWORK

7.1 Introduction....169
7.2 The methodology....169
7.3 Example dialogues and historical approaches....171
7.4 The electricians' negotiation....173
7.4.1 A summary of the principles of cooperative multi-agency under scrutiny....174
7.4.2 The electricians' informal negotiation - utterance analysis....175
7.4.3 The electricians' informal negotiation - conclusions....183
7.5 "Othello", Act III, Scene III....184
7.5.1 The context....185
7.5.2 "Othello" - utterance analysis....189

7.5.3 "Othello" - an alternative analysis.....192
7.5.4 "Othello" - conclusions.....196
7.6 Conclusions.....197

CONCLUDING REMARKS 199

BIBLIOGRAPHY.....203

APPENDIX 1 Transcript of the electricians' informal negotiation 218

APPENDIX 2 Extract from "Othello", Act 111, Scene 111.....219

APPENDIX 3 An example analysis from the electricians' informal negotiation (from M's point of view.....220

INTRODUCTION

Research Aims

This thesis describes research aimed at the development of a theoretical framework for computer models of multi-agent dialogue. It has been designed as a theoretical preliminary to future implementations of cooperative systems which use dialogue to negotiate and resolve differences. These will be equally applicable to human computer interaction (HCI) contexts, as well as distributed artificial intelligence (DAI), the latter concerning purely machine interactions.

Motivations

In social psychology circles, conflict management and conflict resolution are considered to play an important and positive role in cooperation and the maintenance of social stability. These ideas are now firmly embedded in current thinking and research, having been revived in the fifties by Coser (1956), from the classical work of Simmel (Simmel, 1955), and others who wrote at the turn of the century. In artificial intelligence however, most existing research involving aspects of cooperative multi-agent interaction, has assumed that being a cooperative agent means being benevolent; cooperative agents are always in agreement and ready to adopt each others goals. Conflicts either simply never arise, or alternatively they are avoided when they do arise.

This research was motivated firstly by a belief that intelligent machines engaged in joint execution, management, allocation of tasks in the real world, will inevitably be faced with the sort of conflict situations which arise out of a constantly changing and unpredictable environment, just as do human agents. The machinery needs to be available for such automated agents to potentially resolve differences, as opposed to avoiding or ignoring them. Conflict is considered to be a positive force in the maintenance and evolution of cooperative multi-agent systems, because its expression and consequent potential resolution or management makes possible a flexibility in dealing with unexpected events. This is the way that multi-agent systems as a whole can evolve cooperatively, and potentially appropriately to

changing and unpredicted circumstance. In contrast, existing systems are rigid and constrained by imposed benevolence. The second motivation for this research was therefore the belief that the attitudes and experience reflected in thirty years of conflict studies in social psychology regarding conflict's positive role with respect to human agents, is also relevant to the current development of computational models of cooperative multi-agent interaction.

Managing conflict involves a choice of best action, given the conflict which exists. Resolving conflict on the other hand, involves "changing someone's mind", the conflict being thereby removed. Dialogue is therefore the means by which conflict resolution can occur, because dialogue effects changes to agents' mental states. This research concerns conflict resolution; this was the motivation for focussing on the role of dialogue in cooperative sytems. The proposed framework incorporates a pragmatic linguistic theory whereby dialogue comprises utterances, and utterances are speech actions generated and interpreted in context, as attempts to satisfy communicative goals. The context comprises the agent's assumptions about the world, which are represented in her mental states. These include beliefs and goals about the interaction, as well as principles concerning the nature of multi-agenthood and interaction in general. The nature of conflict and cooperation for example, are represented as patterns of mental states. Agents therefore have the means of both recognising and manipulating these states. As autonomous agents however, each only has partial control over this process. Agents are neither assumed to simply benevolently adopt others' communicative goals, and nor are speech actions assumed to be always veracious and open expressions of the speaker's mental states. Each agent has control over what they personally acquire, and what they reveal in dialogue.

Original Contribution

Ideas are incorporated from social psychology and game theory into a new theory of multi-agent interaction. This forms the context within which the pragmatic theory of dialogue adopted from Cohen and Levesque (1987b) operates. It offers a strategic basis for reasoning about the generation and interpretation of speech action, which acknowledges the positive role of conflict in multi-agent cooperation. The resulting theoretical framework models cooperative dialogue between autonomous

agents, each with an element of the control over the information flow between them. They can recognise and alter conflict relationships between them. They can negotiate agreement where there had previously been disagreement.

Definitions of conflict, cooperation and indifference are introduced as propositional postures, which relate specifically to the multi-agent context. Preferences and interests are introduced as properties of single agents, defined in terms of the primary properties of beliefs and goals. These are used to characterise communicative acts as well as the proposed conditions under which autonomous agents adopt beliefs and goals in dialogue as strategic interaction. Veracity, mendacity, concealing and revealing are defined as expressions or properties of speech actions, relevant to the strategic approach.

Theoretical Background

The research described here is an extension of the recent developments in speech act theory of Cohen and Levesque (1987b) and Perrault (1987). Aspects of the work of Rosenschein (1985) and Rosenschein and Genesereth (1985) form the grounding upon which the social psychology theory of Coser (1956) and Simmel (1955) is incorporated into the proposed theory of multi-agent interaction. The properties of communicating agents which relate to strategic interaction owe much to the theories of Schelling (1960), Goffman (1970) and other game theorists, such as Howard (1971). The resulting model of agents and multi-agents is expressed formally using a logic adopted from Cohen and Levesque (1987a, 1987b) which is based upon the adaptation of the possible worlds approach of Kripke (1963) to epistemic logic, by Hintikka (1962).

Methodology and Testing

The methodology has been to iteratively formalise and test the theoretical intuitions. The rationale behind the use of logic to express the theoretical ideas as a preliminary stage to future computer implementation is that the ideas are expressed independently of the mechanisms which would enable

them to be used in a program. This not only avoids machine-oriented technical problems, which detract from the theoretical focus of the research exercise, but also leaves the ideas entirely open to examination and testing. The choice was also made in order to better develop and compare with previous research which has applied formal approaches to the modelling of agents and multi-agent interaction.

The emphasis of the research programme is the development of theory; the specification and description of the problem is the prime concern as opposed to its implementation details. Various problems are acknowledged related to the nature of the particular formal language chosen as the means of expression. It does however, provide a tractable system with a precise semantics to which all expressions must conform. It provides a rigorous base according to which the various paths of reasoning necessary to generate different example dialogue actions can be traced and tested. Sample dialogue actions from conversations from (i) a protocol of a negotiation between an electricians union and their management, and (ii) Othello by Shakespeare, Act 3 Scene 3, have been used for this purpose. Incremental improvement of the theory has occured via the retrospective analysis of these existing dialogue phenomena. They have been used therefore, as a means of evaluating the theory in terms of its explanatory and predictive potential.

Comparisons have been carried out with the approaches of Cohen & Levesque (1987a, 1987b) and Perrault (1987), also via analyses of some of their examples.

Disclaimer

It should be noted that there is no claim of psychological plausibility in this work. This framework is proposed and tested for its validity as a theoretical base from which computational agents can reason about cooperative dialogue. It is as a preliminary stage to machine applications encompassing negotiation style dialogues. It is not intended that conclusions be drawn which relate to mechanisms of human reasoning.

Thesis Outline

Chapter 1 describes the linguistic theory proposed for this framework for computer models of cooperative dialogue. It includes a literature survey reflecting the theoretical background and historical development of speech act theory, in the context of other pragmatic approaches to modelling dialogue in artificial intelligence. The speech act theory adopted considers communication to be grounded in general principles of rational, cooperative interaction. Chapters 2 and 3 therefore describe and discuss firstly the theory of rational agenthood, and secondly the theory of cooperative multi-agent interaction which comprise such principles, for this framework. Some of these are adopted from previous work in artificial intelligence, but others are derived from ideas generated within other disciplines such as philosophy and psychology. Chapter 4 concerns the means by which these principles are expressed for demonstrating, testing and evaluating them as a basis for cooperative dialogue. The background and justification for the chosen formal language is given, followed by the precise model of agenthood proposed. Chapters 5, and 6 then develop the focal issues of the theory. Definitions of conflict, cooperation and indifference, the role of these in cooperative dialogue, and theories of conflict as a positive force in the maintenance and evolution of cooperation, are to be found in chapter 5. Chapter 6 concerns issues related to the control of information in dialogue as strategic interaction. Agents have only partial control over the outcomes of their dialogue actions. They and others are autonomous over the mental states they acquire. This manifests itself in the conditions under which others' beliefs and goals are acquired in dialogue. Secondly, agents are autonomous over their mental states in terms of what is revealed. This property manifests itself in the ability to perform veracious or mendacious, and revealing or concealing expressions. A communicative act is defined in the terms of this strategic framework, and this is contrasted with a comparable definition of Cohen and Levesque's, and their approach in general. The testing and evaluation of the entire framework is then described in chapter 7. Two example dialogues provide instances of dialogue actions which are analysed according to the proposed theoretical principles. Complete transcripts of these dialogues comprise appendices 1 and 2. Some historical insights are also given in chapter 7 into the practical development of the issues and methodology, during the course of the research programme.

Illustrating the Framework

Cohen and Levesque's paper, "Persistence, Intention and Commitment" (Cohen and Levesque, 1987a) began with a little story about a household robot of the future, called Willie. By the end of the paper, it was claimed that Willie could be a robot committed to his goals. This meant that in response to the command: "Willie! Bring me a beer", he should helpfully take on his master's goal, having no existing contradictory one, and assuming that this master (let's call him Fred), was sincere in his communicated desire for a beer. He should be committed to it in the sense that it will only be abandoned once achieved, or if impossible to achieve - there being no beer in the fridge for example, or if the reason for the goal is no longer true. An example of this would be Willie believing that Fred is no longer thirsty.

Three additional scenarios involving Willie and Fred are briefly examined here. The purpose for these is to illustrate those particular aspects of computational modelling of cooperative dialogue which are the focal concerns of this thesis.

Scenario Number 1: Imagine Willie is on his way to the kitchen to get Fred's beer when the voice of Sally, another member of this household, calls out from the bathroom: "Help! Willie, come in here quick!" This is the kind of unexpected event that can occur so often in the real world. The actors, or agents in such a world need to be flexible and able to reassess their goals in the light of new information. Perhaps Sally cannot turn off the taps and there is a danger of flooding. Perhaps she is drowning. The goal she has communicated to Willie may be much more important in terms of their survival, than getting a beer. According to Cohen and Levesque's framework, agents are always helpful in taking on another's goal, but not if they have an existing contradictory goal. Willie simply could not therefore assist Sally at that time because he has Fred's beer to get. In most other existing frameworks for modelling cooperative action, cooperative agents generally do not have conflicting goals; the context is constrained such that this never occurs. Alternatively, conflicts may be acknowledged but avoided in the planning of non-contentious alternatives.

In this framework, Willie is an autonomous agent. On recognition of Sally's goal for him to help her in the bathroom, he neither benevolently adopts her goal because she has it, nor rigidly resists it

because of his existing contradictory goal. He examines his preferences with respect to the alternatives and circumstances with which he is faced. Preferences are based upon maximal satisfaction of goals and consistency of belief, taking into account that goals and beliefs are held with different strengths or centrality to the agent. Acting according to preferences in situations of choice therefore implies greatest consistency with the "values" the agent holds. It is imagined that Willie would be programmed to prefer to help humans in trouble than perform household chores, if such situations should ever arise. He drops Fred's goal and adopts Sally's.

Commitment to this goal requires that it is not already achieved and is not impossible. Willie must also believe Sally really wants his help. Is this a true expression of her desires? He cannot assume she is always sincere. Did she make the request because she wanted his help and believed that Willie would take this goal on and thus respond favourably? If he has beliefs about Sally, for example, which include that this is not the kind of thing she does for a joke, or perhaps there was a sense of urgency in her voice, or....., then he will conclude that it was in fact a legitimate request. Sally understands this and can therefore reason that it was okay to ask Willie to help; it was a good strategy for her. She wanted Willie's help and had expectations regarding his preferences and beliefs, such that not only would he actually adopt her goal, but his subsequent action would also be in her favour.

The justification and motivation for this strategic approach to a theory of dialogue, is in the development of a theoretical framework for the computational modelling of dialogue appropriate to real world, unpredictable multi-agent situations. Agents must be able to behave flexibly in the light of new information. This may involve the resolution of conflicts or differences between them. For example, Willie had a goal to be getting Fred's beer but Sally is faced with a real and unexpected problem. Suddenly Willie is being asked to adopt Sally's contradictory goal. Real environments cannot be constrained to the point where different agents with unchangeable goals can be assumed to be forever in agreement.

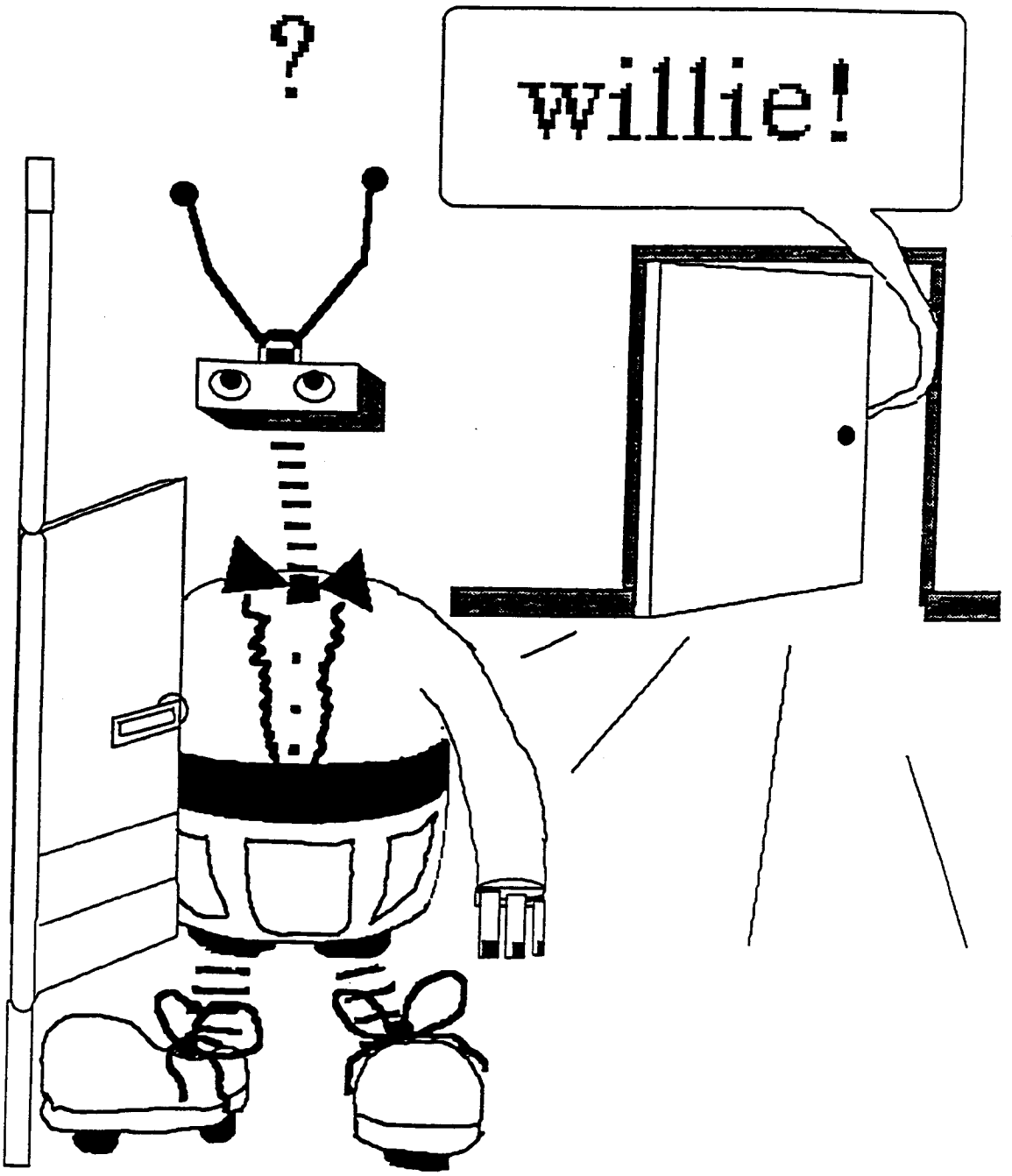
In this framework, conflict is considered as a positive force in the maintenance and stability of social systems. The next two scenarios are aimed at demonstrating that the use of dialogue to express and potentially resolve conflict, offers flexible solutions ensuring the continuation and evolution of the cooperative system as a whole, in a changing and unpredictable environment.

Scenario Number 2: Fred, discovering Willie on his way to the bathroom before he has got the beer, is very cross and demands to know why Willie is being disobedient. There is a conflict between Willie (who wants to help Sally) and Fred (who wants Willie to get his beer). Willie recognises this. He also knows that if Fred and he had a common goal regarding Willie's current activities, they would no longer be in conflict. How to achieve this? Dialogue affects mental states. If he informs Fred of the situation, Fred might then drop his goal for Willie to get the beer because he believes Fred to have a preference for Sally not to have flooded the bathroom than be drinking beer. The conflict would thus be resolved. Alternatively, of course, Fred may subsequently just tell Willie not to be so stupid, and Sally can sort herself out. If Willie accepts this, the conflict is also resolved by Fred having used dialogue to alter Willie's goal such that their goal is common. Whichever of these is the case, cooperation is resumed between Fred and Willie.

The possibility of "changing someone's mind" means flexibility, and flexibility of action is crucial to action being appropriate in changing conditions. This example demonstrates how being able to "change someone's mind" can resolve conflict and maintain cooperation between multi-agents. Both Fred and Willie understand the nature of conflict and cooperation. These are contexts of the multi-agent system, determined by their own mental states. "Changing someone's mind" therefore means altering that context; it can change conflict to cooperation. They also understand the role of dialogue as the manipulator of these mental states.

Scenario Number 3: Willie is in the process of doing the ironing, when he recognises Fred's goal for him to get a beer. There is a conflict between Willie's existing goal to be ironing and Fred's goal for Willie to get the beer instead. Willie has a preference that dictates that he maintain his goal to be ironing in these particular circumstances. He could therefore just ignore Fred and carry on ironing. However, having a representation of the nature of conflict and understanding that he is in that situation, he understands that the conflict is not resolved by this action. Fred will still try to get him to adopt his (i.e. Fred's) goal. Therefore, he and Fred need to reach some agreement about the situation; the conflict needs to be resolved. He knows that conflicts are resolved when one party no longer wants his own goal but adopts the other's. His attempt to use dialogue to create this situation may be something like:

"Well, if I have to go and get a beer now, you'll have to iron your own shirt!"



CHAPTER 1: A pragmatic theory of dialogue

1.1 Introduction

The term "dialogue" is defined in the Oxford English dictionary as: "A conversation or verbal exchange of thought, between two or more persons". Conversation is defined as "the action of consorting or having dealings with others", and "exchange of thought" is part of the definition of communication (English Oxford Dictionary). The aim of eventually achieving a computational model of dialogue is therefore one where the objective is an automated system comprising verbal communication between multi-agents. There are various approaches to the study of verbal communication or language usage, also termed "pragmatics", which have been applied within artificial intelligence. A description of the overall artificial intelligence (AI) perspective and justification for the particular approach adopted in this framework, is given in section 1.2.

The term "pragmatics" was initially distinguished from the syntax and semantics of signs by Morris (1938). He defined pragmatics as the study of the relation between signs and their interpreters or users. This contrasts it from semantics which was defined as the study of the relation between signs and their denotations. Semantics is about the sentence meaning inherent in every utterance of a particular sentence. However, the same sentence can be uttered on different occasions with a variety of intended meanings. Pragmatics concerns utterances and these have both linguistic and non-linguistic properties such as, the particular speaker's intentions in making the utterance, its time and place, and so on. It concerns the use of language.

> "Pragmatics is the study of the relations between language and context that are basic to an account of language understanding" (Levinson, 1983).

This research adopts a pragmatic theory of dialogue. Utterances are speech actions which effect changes to the multi-agent system; they effect changes to the mental states of the interacting agents. They are interpreted and generated in the context of those mental states. These comprise importantly the

communicative goal of the speaker. In addition however, they must comprise a general understanding of the relation between such communicative goals and rational action in a multi-agent environment. This outline of a theory of dialogue describes the recent developments in speech act theory, as proposed by Cohen and Levesque (1987a, 1987b) and Perrault (1987). Precisely what these recent developments are, their theoretical background and historical development is explained and discussed in section 1.3. The role they play in the proposed framework for dialogue is described in section 1.4.

> "The set of premises used in interpreting an utterance (apart from the premise that the utterance in question has been produced) constitutes what is generally known as the *context*. A context is a psychological construct, a subset of the hearer's assumptions about the world. It is these assumptions, of course, rather than the actual state of the world, that affect the interpretation of an utterance. A context in this sense is not limited to information about the immediate physical environment or the immediate preceding utterances: expectations about the future, scientific hypotheses or religious beliefs, anecdotal memories, general cultural assumptions, beliefs about the mental state of the speaker, may all play a role in interpretation" (Sperber & Wilson, 1986).

1.2 Approaches to modelling dialogue in artificial intelligence

1.2.1 Aims and applications

Research into the computational modelling of dialogue has been aimed largely at achieving dialogue or discourse between humans and machines - human computer interaction or HCI. Cooperative ventures necessitate communication, and the ultimate goal is to eventually achieve natural language interfaces for expert systems, database packages, educational programs and so on, which do not require the user to have knowledge of specialised and restrictive command languages. Other applications include text understanding and machine translation. In addition, cooperative planning and task execution between machines on distributed networks in distributed artificial intelligence or DAI, also requires an understanding of the principles of communication, although not necessarily within the confines of

natural language.

1.2.2 Methods and theories

The various approaches to modelling dialogue in artificial intelligence and computational linguistics can be broadly categorised as either psychological (intentional), or structural approaches to language processing. The former are various interpretations of the work of Grice (Grice 1957, 1969) , which consider the recognition of the speaker's intentions in producing an utterance to be a crucial component to understanding its meaning. This is a fundamental component of Searle's speech act theory (1969, 1975). In fact, the research work into computational modelling which has taken place to date which focuses on the intentional component of utterance interpretation and generation, has adopted (varying interpretations of) speech act theory, as its theoretical foundations. Examples are the work of Cohen (1978), Cohen and Perrault (1979), Allen & Perrault (1980) , Appelt (1982, 1985), Cohen and Levesque (1987b), and Perrault (1987).

Structural approaches on the other hand, focus on analysis of structural relations between elements of discourse, discourse being language behaviour which typically involves multiple utterances as well as multiple participants. Reichman (1985) for example, has devised a set of discourse processing rules specified in terms of an abstract grammar of discourse processing and without access to individual speaker beliefs or knowledge structures (Reichman, 1985). Grosz and Linde studied task-oriented dialogues which they parsed according to distinct structural discourse elements (Grosz, 1981, Linde,1974). Power described conversational procedures, which are lists of instructions as to how utterances in pairs should be produced and interpreted (Power, 1979). His work followed Schegloff and Sacks observations concerning adjacency pairs in conversational structure (Schegloff & Sacks, 1973). ...and there are many more.

Mostly, these approaches are not considered as mutually exclusive alternatives, but two aspects of the problem which future practical systems will need to embrace. The recent work of Grosz and Sidner demonstrates this by suggesting that discourse structure has three components which deal with different aspects of the individual utterances: the linguistic structure, the intentional structure, and the attentional

state (Grosz & Sidner, 1986). They acknowledge both the central role of intentions, which they differentiate into discourse purposes and discourse segment purposes, and their dependency upon an adequate theory of intention and action. Reichman (1985) however, is one exception to this. She considers that identification of the structural units conversational moves, is sufficient to understanding, and these are derivable solely from the discourse context and conventional rules of discourse processing. Her research aimed to formalise the repertoire of constructs relevent to conventional uses of language. Even if the speaker's intentions are considered relevent, they are conventionally communicated and correlated to conversational moves (Reichman, 1985).

1.2.3 Speech as planned action

Both structural and intentional approaches generally encompass an artificial intelligence perspective. This considers language processing as problem solving behaviour. Agents have a task, or problem to solve. They generate plans to break down the task into achievable sub-tasks. Language is a means of potentially achieving these, and the utterances produced reflect the plan. Planning therefore is an essential element in language production and understanding.

Traditionally, planning systems in artificial intelligence such as STRIPS (Fikes and Nillson, 1971) and NOAH (Sacerdoti, 1977), consider actions as operators, defined in terms of the preconditions which must be true for the action to take place, the effects which are obtained when the action has been performed, and the body of the act, which describes the means by which the effects are achieved. The plan is a sequence of actions which transform the initial world state where certain preconditions are true, into a goal state where the desired effects are successfully attained. Plans can be generated to successfully achieve some task (Grosz ,1981, Allen & Perrault, 1978, Litman, 1983). Alternatively, they may be designed to maintain certain goals of the discourse itself, such as coherence (Clark & Marshall, 1978), being social, projecting an image (Hobbs, 1980).

The main distinction between the structural and psychological approaches is whether interpretation by plan recognition is aimed at determining the speaker's various goals and whether this should be according to specified relationships between agents beliefs, goals and action, or whether it is the

structures of conversation which are of prime importance. Historically, both approaches have focussed on the initial recognition (and correspondingly, generation) of discernable conversational structures. These could be speech acts (Allen, 1980, Cohen & Perrault, 1979, Searle, 1979), scripts and schemas (Schank & Abelson,1977, Dyer, 1982), dialogue games (Levin and Moore, 1977) conversational procedures (Power, 1979), conversational moves (Reichman, 1985).....and so on. All participants have representations of these units for recognition and via which actions can be deemed appropriate for generation. Speech acts are characterised in terms of the agent's beliefs and goals however, whereas the alternatives comprise specified sequences of multiple sub-actions in the form of schemata or stored plans. It is generally acknowledged that the problem with any approach requiring the recognition of specified and characterised structures, is that it demands their orthodox and standard use in conversation. All language structures must be employed in ways for which suitable representations currently exist within the stored "library" of structural categories. Natural language useage is simply not that constrained.

The theory of dialogue adopted by this research is that suggested in the recent work of Cohen and Levesque (1987b). It is a re-interpretation of the psychological/intentional approach encompassed in speech act theory which avoids the constraints referred to above. This is because the focus is shifted away from the structures or acts themselves, and onto the mental states, or beliefs and goals which underlie the performance of acts. Interpretation concerns the recognition of these mental states, which is based upon an understanding of the general principles relating mental states and speech action. The emphasis is therefore not on the act but on the speaker's goal that the hearer adopt a particular mental state, inferred from an understanding of rational interaction and contextual beliefs. The requirement that a particular speech action be recognised first as a request or an assertion for example, in order to then determine intention, is now considered redundant (Cohen & Levesque, 1987b, Perrault, 1987).

Section 1.4.5 includes a more complete explanation of Cohen and Levesque's (1987b) and Perrault's (1987) recent developments in speech act theory. It requires however, an initial discussion of the general theoretical background and historical development of speech act theory, as well as a brief review of some of the earlier attempts at computational application.

1.3. Speech act theory

This section offers a critical and historical approach to speech act theory, in order to introduce and justify this framework's linguistic background.

1.3.1 Theoretical background and historical development

The view of speech as action planned to achieve goals stems from the work of Austin (1962), Grice (1957, 1969) and Searle (1969). The starting point was the work of Austin (1962). His major insight was that utterances do more than express things about the world. They actually change the state of the world. Changes are effected by actions, and therefore uttering a sentence is considered as the performance of a speech action, to be distinguished from the truth conditions of propositions contained in the sentence. Austin's work has parallels with the earlier work of Wittgenstein. In "Philosophical Investigations", 1958, Wittgenstein suggested that utterances should be explained in terms of the purposes for which the language is being used.

Austin analysed performatives, demonstrating that speech acts such as promising, warning, declaring, are not just true or false, but successful or unsuccessful (felicitous, or unfelicitous). Such acts are performed with the purpose of their felicitous achievement. Utterances accomplish actions by having certain forces. These are locutionary force, which comprises the content of what is said, the illocutionary force, which comprises the intent or conventional force of what is said, and the perlocutionary force which is the effect of what is said. In performing one sentence therefore, Austin's suggestion was that three types of speech act are being performed. Firstly, the locutionary act which comprises words which satisfy the vocabulary and grammar of the language. These are used in the performance of the illocutionary act. Examples are a statement, a request, or a warning. These contain the propositional content specifying what is being stated or requested or warned about. Thirdly, there is the perlocutionary act which is the act performed by making the utterance. An example of this is the hearer being warned.

Austin's main interest was the illocutionary act and this is now considered synonomous with the

term "speech act". The distinction between the illocutionary and perlocutionary acts is very vague. They differ in that illocutionary acts can be achieved by the conventional performance of an utterance, and yet the perlocutionary act is dependant on circumstance. Perlocutionary acts can be said to concern the consequences of what is said and yet some illocutionary acts such as promising, have built-in consequences (Levinson, 1983). In addition, an illocutionary act is successful if it results in "securing of uptake" by the hearer. This means that the force and content of the utterance is recognised. However, securing of perlocutionary effects is beyond the speaker's control. The hearer may successfully recognise the speaker's act as an assertion, but remain unconvinced.

Searle developed Austin's work on speech acts by attempting to outline the necessary and sufficient conditions for successful performance of illocutionary acts (now simply referred to as speech acts) (Searle, 1979). This was as a result of believing that speaking a language is to be engaging in rule-governed behaviour, of which there are two types: regulative and constitutive. The former regulate behaviour, such as rules of etiquette; the latter constitute the activity itself. These are the rules of the game. What are these constitutive rules, or conventions for successfully using language?

Some conditions are general to all kinds of speech act. For example, both the speaker and hearer comprehend the utterance, or they are both conscious, and so on. However, others will be specific to the particular act in question. Searle devised a classification of types of speech acts, in order to compare these in terms of the necessary and sufficient conditions for the successful performance of each. These were termed felicity conditions. First the classification:

Utterances are *directives* whereby the speaker wants the hearer to do something. Examples are a request or a command; or

representatives which relate to the truth of the proposition expressed. Examples are to assert or lie or conclude; or

commissives , such as a promise, threat or offer; or

expressives which express a psychological state. Examples are to thank, apologise, congratulate or welcome; or

declarations which change the state of the world. Examples are christenings, declaring war or excommunicating.

From the comparison of necessary and sufficient conditions for the effective performance of these, Searle extracted "sets of semantic rules". These comprise physical enabling conditions as well as an intentional component. The following is an example of the conditions for a request:

S performed a request R that p in uttering x to hearer H, if the following conditions hold:

Input/Output conditions - S and H are paying attention

Propositional content conditions - p is about a future action of H

Preparatory conditions - H is able to do an action which will bring about p; S believes this; it is not obvious to S that H would do this act to bring about p in the normal course of events

Sincerity condition - S actually wants H to do the action which will bring about p

Essential condition - S intends that the utterance x should count as an attempt to get H to do the above action

Gricean condition - S intends to produce in H the belief that the essential condition holds

Semantical condition - The semantical rules of the language are such that x is correctly and sincerely uttered iff all the above conditions hold.

Satisfaction of the intentional conditions in the above necessitates recognition of the utterance's illocutionary force and propositional content. H has knowledge of the conventions of the language and hence recognises the indicator of the type of speech act being performed. This is the illocutionary force indicator. H also recognises the relation between this and the utterance's propositional content.

Searle's intentional conditions are derived from Grice's notion of $meaning_{nn}$ or non-natural meaning (Grice ,1957 & 1969), which can be summarised as follows:

In order for an utterer U, to convey meaning to an audience A, by an action or utterance x,

(i) U thinks that x can induce a response (mutual knowledge of the "crucial features" of utterances and their correlation by convention or association with the desired response, is assumed)

(ii) U intends to induce that response

(iii) A recognises U's intention

(iv) A recognises U's intention that the intention be recognised. (This is a reflexive intention, which is an intention that is intended to be recognised as intended to be recognised.)

Non-natural meaning is distinguished from natural meaning, which is "directly and openly telling someone something" by virtue of the response being at least in part, on the basis of this recognition of intention. In addition, the audience is required to infer something - "getting someone to think something" (Grice, 1957).

There have been several papers and articles written which are critical of some aspect of the Gricean position (Allwood, 1976, Hare, 1967, Mackay, 1973, Schiffer, 1972, Wilson, 1970, Wright, 1975, Ziff, 1967). Differences include: how recognition of intention takes place, the necessity for reflexive intention, and the role of convention. Searle's account is as follows:

> "In speaking I attempt to communicate certain things to my hearer by getting him to recognize my intention to communicate those things. I achieve the intended effect on the hearer by getting him to recognize my intention to achieve that effect."....."in virtue of his knowledge of the rules for the sentence uttered" (Searle,1969).

This does not include Grice's reflexive intention. Instead, the inclusion of illocutionary force recognition in Searle's speech act theory implies that the response is due to recognition of intention, but via recognition of a particular conventional type of act. The major difficulty this semantic approach poses is with the interpretation of non-literal or indirect speech acts. Direct speech acts are those where the syntactic form of the utterance corresponds directly to the illocutionary force of the utterance. An example is an imperative utterance being used to issue a command , such as "Open the door!". Indirect speech acts however, do not correspond directly to intended illocutionary force. "Would you mind opening the door?" for example, is similarly a request for the door to be opened. Gordon and Lakoff (1975) proposed dealing with the problem of indirect speech acts by devising a set of conversational postulates by which the literal form of one speech act could be said to "entail", or be directly related to the indirect form of another. Searle however, rejected this approach and proposed the following :

"In indirect speech acts the speaker communicates to the hearer more than he actually says by way of relying on their mutually shared background information, both linguistic and non-linguistic, together with general powers of rationality and inference on the part of the hearer. To be more specific, the apparatus necessary to explain the indirect part of indirect speech acts includes a theory of speech acts, certain principles of cooperative conversation (some of which have been discussed by Grice (1975)), and mutually shared factual background information of the speaker and the hearer, together with an ability on the part of the hearer to make inferences" (Searle, 1975).

Recognition of indirect speech acts is related to the interpretation of implicatures. Implicatures are non-conventional inferences, intentionally conveyed. "The battery's gone flat" could be an explicit assertion, for example. It could have been said with the intention that it be interpreted, and therefore assigned to speech act types as either an accusation that the hearer shouldn't have let the battery go flat, or perhaps an order that the hearer should get the battery recharged (Sperber & Wilson, 1986). The theory of understanding such conversational implicatures, referred to in the quote from Searle above, involves Grice's cooperative principle (Grice, 1975). This comprises four maxims of efficient, rational, cooperative conversation:

(i) *quality* - don't say what you believe to be false or that for which you lack evidence

(ii)*quantity* - make your contribution no more or less informative than is required

(iii) *relevance* - make your contributions relevant

(iv) *manner* - avoid obscurity, ambiguity, be brief, be orderly.

Grice claims that we do not necessarily adhere to these principles at a superficial level, but if we interpret what others say as conforming to them at some level, then we can make inferences in addition to those from solely semantic content. For example:

A: "Where's Bill?"

B: "There's a yellow VW outside Sue's house"

B's reply apparently violates the maxims of relevance and quantity. However, if it is assumed that B is adhering to the maxims at a deeper level, then the implication is that the VW must belong to Bill and thus B is communicating to A that Bill is at Sue's house.

Maxims can be deliberately flouted as is the case with tautologies such as "Boys will be boys" or "Either John will come or he won't". These violate the maxim of quantity. Assuming the cooperative principle and given the particular context of the utterance, appropriate inferences can be made.

> "Speech act theory thus offers itself as a natural complement to Gricean pragmatics, dealing with the classification in speech act terms of both explicatures and implicatures" (Sperber & Wilson,1986).

Searle's speech act theory (1975) was developed by Perrault & Allen (1980). The significance of this work (which is described in section section 1.3.2.1) is that it laid the theoretical groundwork for research into formalising speech act theory for computational applications. The early work in this area was carried out by Cohen and Perrault (1979), Allen & Perrault (1980) and Appelt (1982, 1985).

1.3.2 Formalising speech act theory for computational models - the early work

Perrault & Allen (1980) introduced an intermediary level between surface form conditions and illocutionary acts, called the surface acts. Surface level acts are realized literally and then recognised as having been performed with the intention that the hearer infer (using plan inference rules and associated heuristics) that the speaker wants to achieve the effects of a particular illocutionary act.

> "A speaker can perform one speech act A, by performing another speech act B, if he intends that the hearer recognise not only that B was performed but also that through cooperative behaviour by the hearer, intended by the speaker, the effects of A should be achieved." (Perrault & Allen, 1980).

Their hypotheses were based on an understanding of the participants in the dialogue as rational

agents; rational agents which have beliefs and goals and engage in goal seeking behaviour, amongst which is the modification of the beliefs and goals of other agents. They can identify the actions, and from these the goals, of others. They are also capable of cooperative behaviour, described as adopting another's goal and attempting to achieve it[1]. Agents achieve goals by plan construction and then execution, the plan being one which changes the current state of the world into one in which the goal obtains. If this rationality is imputed to others, then it can be assumed that each agent constructs and infers plans similarly. Therefore, on the basis of another's observed action, a (possibly empty) set of expected goals, and some rules of plan inference and construction, partial plans can be constructed. Heuristics for rating these suggest the most likely plan that the speaker is executing.

A model based on the above theory and for application in a natural language system, is described in Allen & Perrault (1980). It was implemented in the domain of train times, and tested in terms of providing mechanisms for analysis of indirect speech acts. It was also designed to generate responses which provide more information than required. An example of this:

patron : "When does the Montreal train leave?"

clerk : "3.15 at gate 7."

In this example, the clerk provides information as to the departure location, believing that the patron has goals such as meeting or boarding the Montreal train. This goal is believed by the clerk to be an obstacle. Obstacles are goals which cannot be achieved without assistance. He adopts this goal as his own and plans to achieve it. Execution of the plan is his response.

Speech acts are modelled in the system as parameterised procedures, the parameters being a speaker, a hearer, and a propositional content. The execution of a speech act leads to the production of an utterance. They are described in terms of preconditions and effects, these being defined according to the speaker and hearer's beliefs and wants and independently of syntactic form. For example:

[1]Theories of rational agenthood and cooperative interaction are discussed in Chapters 2 and 3.

INFORM (speaker hearer, prop)

precondition: speaker KNOW prop

effect: hearer KNOW prop

body: hearer BELIEVE speaker WANT hearer KNOW prop.

Surface speech acts which correspond directly to the form of the utterance are used to handle the problem of indirect speech acts, as described at the beginning of this section. An indicative mood sentence is always an S.INFORM act, for example. The effect of an INFORM matches the body of an S.INFORM for direct interpretation, but there are other surface acts which also generate the effects of an INFORM.

S.INFORM (speaker hearer, prop)

body: hearer BELIEVE speaker WANT hearer KNOW prop.

Belief representation is according to the Hintikka schemata which provides beliefs with certain properties. This is explained in detail in chapter 4. Wants are also represented, but in this work the properties of WANT are specified merely by the planning and plan inference rules. A plan is a sequence of actions which transforms an initial world state, W into a goal state G. Plan construction is accomplished by finding a sequence of actions which will accomplish the transformation from W to G. This is done by backward chaining. Given a goal G, what action A has G as one of its effects? If the preconditions of A are not satisfied, then they become sub-goals and the process is repeated. The bindings for the parameters of the actions in the constructed plan then have to be specified. The planning process is characterised as a set of planning rules and a control strategy. An example of a planning rule is as follows: if A wants to achieve X then he may want to achieve Y.

Plan inference takes observed actions and attempts reconstruction of the speaker's plan, using knowledge of the way plans are constructed, and beliefs about the acting agent's possible goals. Again there is a set of inference rules and a control strategy. An example of an inference rule: if S believes A has a goal of executing ACT, and ACT has an effect E, then S may believe that A has a goal of achieving E. Partial plans are created, some constructed as alternatives from the observed action by plan

inference, and others called expectations are constructed using plan construction rules on expected goals. These are then rated according to probability of being the correct plan, which is determined using a set of heuristics.

Allen and Perrault's speech act definitions were developments of Cohen and Perrault's earlier work in which speech acts such as requesting and informing were suggested as being modelled as planning operators, defined in terms of the speakers' and hearers' beliefs and goals (Cohen & Perrault, 1979). The following example shows their definitions to only have had preconditions and effects:

INFORM (speaker, hearer, prop)

CANDO PRecondition: prop

WANT PRecondition: speaker BELIEVE speaker WANT inform-instance

EFFECT : hearer BELIEVE speaker BELIEVE prop

Cohen implemented these within OSCAR which was a system designed to inform or request according to a plan for a hearer to recognise the intention to perform a speech act (Cohen, 1978).

Appelt's language generation system incorporated some of the relevant and significant contributions to date in the fields of planning, psychological/ intentional approaches to language generation and reasoning about beliefs, goals and actions (Appelt, 1982, 1985). It comprised an utterance planning system named KAMP, the linguistic component of which was based on Searle's speech act theory. The theory he adopted assumed:

"the speaker intends to achieve a goal that he reasons can be brought about by the perlocutionary effects of a particular illocutionary act performed in the current context. The speaker then plans a surface linguistic act with the right force and propositional content. The hearer recognises the surface speech act by knowing the propositional content and illocutionary force of the utterance, and infers what illocutionary act the speaker intended to perform. From the hearer's knowledge of the conventions governing illocutionary acts, the mutual knowledge he shared with the speaker, and his knowledge of the speaker's intentions, he changes his beliefs or intentions. Ideally these changes

will correspond to the perlocutionary effects for which the speaker originally planned the action" (Appelt, 1985).

The effects of the speech act were therefore realised by the hearer's recognition of the act, this being characterised as intentional. Illocutionary acts were axiomatised in the manner of Cohen and Perrault (1979) and also three surface speech acts, COMMAND, ASK and ASSERT. However, these were only used for the planning of direct speech acts.

1.3.3 What was wrong with the early work?

Current opinion regarding speech act theory now considers the recognition of types of acts as unnecessary to the understanding of utterances.

"Although one can *label* parts of a discourse with names of illocutionary acts, illocutionary labelling does not constitute an explanation of a dialogue" (Cohen & Levesque, 1987b).

"It is one thing to invent, for one's own theoretical purposes, a set of categories to use in classifying the utterances of native speakers, or to try to discover the set of categories that native speakers use in classifying their own utterances. It is quite another to claim that such a classification plays a necessary role in communication and comprehension. To see one type of investigation as necessarily shedding light on the other is rather like moving from the observation that tennis players can generally classify strokes as volleys, lobs, approach shots, cross-court backhands and so on, to the conclusion that they are unable to perform or return a stroke without necessarily classifying it" (Sperber & Wilson, 1986).

In other words, recovering precisely which act has been performed is unnecessary to its comprehension. Speech acts such as bidding in bridge or declaring war are cited by Sperber and Wilson as possible exceptions to this, although they suggest these possibly belong to the study of institutions and not to

the study of verbal communication (Sperber & Wilson, 1986).

Speech act theory has been challenged also in terms of the difficulty of illocutionary act recognition, regardless of whether this is actually a necessary step to understanding or not. Levinson suggests that most language use is in fact indirect, where there is no simple correspondance between utterance form and utterance force (Levinson, 1983). However, he claims speech act theorists are committed to the literal force hypothesis, or LFH, which considers illocutionary force as built into sentence form. Gazdar (1979) summarises this as two rules: Firstly, the force of explicit performatives is in the performative verb in the utterance's main clause. Secondly, the three main sentence types, imperatives, declaratives and interrogatives have associated conventional illocutionary forces of ordering or requesting, stating and questioning respectively. All other speech acts have an additional, inferred force and are therefore indirect. The problem is that there is enormous diversity of actual language useage. An example of this is the fact that the imperative is only rarely used in requesting.

> "On the face of it, what people *do* with sentences seems quite unrestricted by the surface form (i.e. sentence-type) of the sentences uttered" (Levinson, 1983).

Levinson therefore questions the LFH; sentences do not have literal forces. The problem is a general one of "mapping speech act force onto sentences in context" (Levinson, 1983).

Haslett offers further objections to speech act theory. She refers to its "neglect of the listener's role in interaction" (Haslett, 1987) and ignoring of the negotiation of meaning between speaker and hearer. She cites Edmondson's arguments that the hearer's uptake is central to the illocutionary force of an utterance as well as the problems caused by utterances which can be interpreted according to a variety of possible intentions (Edmondson,1981). Finally she agrees with Levinson (1983) that the importance of background and commonsense knowledge possessed by both speakers and listeners is largely ignored in speech act theory.

In addition, adoption of the traditional notion of plans as "recipes-for- action" is currently being challenged. Allen's system as well as Appelt's (Allen & Perrault, 1980, Appelt, 1982, ,1985) construct and infer plans "from a library of similar recipes that are assumed to be mutually known to the actor and the inferring agent" (Pollack, 1987). This is the traditional Artificial Intelligence approach which

Suchman's recent work criticises (Suchman, 1987). Suchman prefers to consider action as situated or ad hoc and circumstancial.

> "every instance of meaningful action must be accounted for separately, with respect to specific, local, contingent determinants of significance" (Suchman, 1987).

Martha Pollack (1987) is not a critic of the planning paradigm, but considers the issue of invalid plans. What if an actor A is relying on a plan which is not in the system's operator library? Her suggestion is that plan inference supporting a theory of cooperative communication must be concerned with "the structure of the complex mental attitude of having a plan, as well as with the structure of the objects of that attitude" (Pollack, 1987). The process of plan inference is then one of attributing a collection of beliefs and intentions to an actor; not only those to believe she has a plan, but also beliefs that explain those beliefs and intentions - the eplan.

1.3.4 Alternative approaches/ theories

Some alternatives to speech act theory as it was realised in the early formal models, are discussed in this section.

Levinson (1983) considers it best either to avoid speech acts altogether in studies of language use based on communicative intention, utterance function and interactive context, or to adopt more of a pragmatic approach, such as context-change theory. The idea in this is that speech actions do more than express meaning, they change the set of background assumptions. Speech acts can therefore be viewed and easily formalised as operations in the set-theoretic sense, on contexts. They act as functions from contexts into contexts (Gazdar, 1981)

Another more pragmatic approach to speech act theory is the inferential theory of Bach and Harnish (1979). It contrasts with Searle's version of speech act theory in considering that for both indirect and literal utterances there is a connection between surface linguistic form and speech acts, but this is not a semantic one; it is inferential. The three factors influencing this are content, context and communicative

intention.

> "Our view is that linguistic communication essentially involves the speaker's having a special sort of intention (an intention that the hearer make a certain sort of inference) and the hearer's actually making that inference" (Bach & Harnish, 1979).

They do however still have a taxonomy of illocutionary acts but these are distinguished by the attitudes the speaker expresses in performing them; attitudes towards the propositional content and the intention that the hearer have or form a corresponding attitude. Communicative illocutionary acts (as opposed to conventional ones, such as voting, resigning, christening etc.) are of the following types:

constatives - express the speaker's belief and intention that the hearer form a like belief. Examples are assertives, predictives, descriptives, or

directives - express the speaker's attitude towards a prospective action of the hearer and the intention that the attitudes expressed by the utterance be taken as a reason for this action. Examples are requestives, questions, or

commisives - expresses the speaker's intention and belief that the utterance obligates him to do something. Examples are promises, offers, or

acknowledgements - express feelings regarding the hearer. Examples are apologising, condoling.

Expressing an attitude via an utterance is reflexively intending the hearer to believe that one has the attitude, because the utterance was made. The fulfillment of the intention is purely in its recognition. This occurs on the basis of what is said, in accordance with two mutual beliefs shared amongst members of the linguistic community. These are a linguistic presumption and a communicative presumption. There are also mutual contextual beliefs.

Sperber and Wilson (1986) also promote an inferential approach to utterance interpretation, but not via speech acts. They consider Searle's model of speech acts "reduces Grice's analysis to a commonsense amendment of the code model" (Sperber & Wilson, 1986). Code models are those where communication is achieved by the encoding and decoding of messages. Coding implies conforming to rules, and

messages are then decoded according to a knowledge of those rules.

> "Grice's greatest originality was not to suggest that human communication involves the recognition of intentions. That much, as already pointed out, is common sense. It was to suggest that this characterisation is sufficient: as long as there is some way of recognising the communicator's intentions, then communication is possible" (Sperber & Wilson, 1986).

They describe an example where Peter asks Mary "How are you feeling today?" Her response is to pull a bottle of aspirin out of her bag and show it to him. Although there is no code or convention which she is following, she is enabling Peter to recognise that she intends him to believe she is unwell. Communication is achieved by the communicator providing evidence from which the audience can infer her intentions (Sperber and Wilson, 1986). The suggestion is that there are two different modes of communication; a coding-decoding mode and an inferential mode. Their work is an attempt to uncover how shared information from which inferences of the communicator's informative intention can occur, is exploited in communication; what is relevance and how it is achieved.

Cohen and Levesque's theory, described in the next section, and providing the basis from which this research develops, is yet another inferential approach, and one incorporating speech acts. It is an adaptation whereby recognition of illocutionary acts is considered unnecessary because "all the inferential power of the recognition of illocutionary acts was already available from other sources" (Cohen & Levesque, 1987b); these other sources being a theory of rational interaction. Allwood similarly considers Searle's conditions to be derivable from those "other sources" , these being principles of rational agenthood, action and cooperative interaction. His theory is that full-blown communication, or "the type of communication which is paradigmatic for normal linguistic interaction of communication" (Allwood, 1976) is a form of ideal cooperation. From an understanding of his principles of normal, rational agenthood (described in Chapter 2), the nature of ideal cooperation, and knowledge of conventions appropriate to communication, he defines communication as "a species of cooperation connected with rather special intentions and purposes" (Allwood, 1976).

1.3.5 Formalising speech act theory for computational models - recent developments from Cohen and Levesque, and Perrault

In the recent theories of communication of Cohen and Levesque (1987b) and Perrault (1987), utterances are considered as special cases of events that change the state of the world; they change the mental states of speakers and hearers. Utterance events are performed in order to effect those changes, and they do so because they signal that the speaker is in a certain mental state, this including an intention that the hearer adopt a mental state.

The advance from earlier work is primarily that the theory of dialogue is in terms of mental states - how mental states lead to action, how those actions affect mental states. Illocutionary acts have a secondary role. In order to demonstrate the redundancy of the illocutionary level, Cohen and Levesque adopt an alternative analysis of action from that typical of Searle's speech act theory. Actions are therefore not seen as composite of interrelated and simultaneously performed actions, such as Gavrilo Princip's simultaneously pulling a trigger, firing a gun, killing Archduke Ferdinand and starting World War 1 (Searle, 1983). Instead, there is a conceptual distinction between events and descriptions of instances of those events, executed in certain contexts, and having various situation-specific effects. Using this analysis, Princip's finger moving is the primitive event, the context is the finger being in contact with the trigger of a loaded gun and the effects are the gun firing, the death and the war. These three aspects comprise the foundational stratum. On top of this is the stratum of descriptions denoting various events, contexts and effects, all interrelated according to the properties of the underlying events. With a suitable logic of events, the properties of the description should be derivable from the properties of an event, context and effects.

In this way, Cohen and Levesque suggest that from a specific utterance event and the context of the speaker's and hearer's mental states, the illocutionary acts performed and the relationships between them are derivable. If it is possible to derive these from first principles, then the power of description is available, yet explicitly recognising this illocutionary level is unnecessary. Their paper formally derives the definition of a request from principles of rational interaction in order to make this point (Cohen & Levesque, 1987b). Appelt questions this aspect of Cohen and Levesque's work. He considers illocutionary acts still to be useful for planning because they provide a convenient level of abstraction

for the planner to reason at with respect to utterance generation, without having to actually construct a surface utterance. He likens this to the use of macro operators for formulating plans in STRIPS (Fikes and Nilsson, 1971) (Appelt, 1985). He also considers that their theory does not take into account that how particular acts are performed at the surface level can be very important to achieving social goals such as politeness, for example. The hearer should be able to judge which act is being performed and if it is being performed appropriately (Appelt, 1985). Cohen and Levesque (1985) acknowledge that although illocutionary act recognition may be unnecessary, it can be still be practically useful. However, the theoretical basis for communication is considered to be the application of general principles governing mental states, to enable inference about the relationships between different types of effects of actions under the conditions of particular mental states of both speaker and hearer. Such general principles embody a characterisation of rational agenthood and cooperative interaction which is independent of theories of speech acts and communication[1]. If a particular syntactic feature in a certain context is recognised, further consequences can be inferred, from those characterised as correlated with the recognised feature. Those consequences are independent from the specific features of the utterance. The speaker is seen to be trying to bring about a particular chain of consequences from one event in context.

Cohen and Levesque consider that their theory of speech acts satisfies the following requirements: utterance form is differentiated from illocutionary force; the major kinds of illocutionary acts, insincere performances, indirect speech acts, self-defeating speech acts such as "I hereby lie that it is raining", acts performed by multiple utterances and multiple acts performed by one utterance - all are catered for.

Perrault's contribution (Perrault, 1987) is to suggest an alternative means of characterising the utterance effects, this being default logic. His claim is that the mental states of the speaker and hearer after an utterance, are strongly dependent on their mental state before. If many of the utterance effects are defined as defaults, then they can be assumed to hold as long as there is no evidence to the contrary. As an example of the advantages of this approach, he contrasts it with Cohen and Levesque's characterisation of the effects of uttering a sentence with the recognisable syntactic feature of its dominant clause being a declarative:

[1]Details of Cohen and Levesque's model of agents are given in chapter 2

Cohen and Levesque's axiom which follows can be explained as: if it is mutually known between S and H that e is an event of utterance of the sentence s, to H, that S is the agent of e, and that s is a declarative sentence with propositional content p, then after the utterance, H believes it is mutually believed between S and H that S intends H to recognise his intention that H believes that S believes that p is true.

$MK_{S,H}$ (UTTER(H,s,e) & AGT (S,e) & ATTEND (H,S) & declarative (s,p)

... after (e,$BMB_{H,S}G_SB_HG_SB_HB_Sp$))

According to the theory, the event and context are S uttering a declarative sentence with propositional content p, whilst S and H are attending. The effect is the mutual belief resulting from that event. A second set of axioms concerning principles of rational interaction are required for the inference that the hearer will also therefore believe p.

If the speaker is lying and this is undetected, all the conditions of the consequent of the axiom and the further consequence, B_Hp can still hold. If the speaker is lying and the hearer knows this, then the consequence inferred is not B_Hp. However, G_SB_Hp ,and $G_SB_HB_Sp$ still hold, although $\neg B_Sp$. With irony, none of the consequent conditions will hold; neither S nor H believe p, and S doesn't intend H to recognise any intention for H to believe p. Perrault's claim is that the predictions of Cohen and Levesque's declarative axiom are therefore too strong. His solution is that the mental states of the speaker and hearer before the utterance be included in such a characterisation.

Utterance effects, such as beliefs, are defined as defaults in Perrault's theory (Perrault, 1987). Those which exist prior to the utterance are assumed to persist. Perrault's persistence theory of belief states that old beliefs persist over time and new ones are only adopted if they are not contradictory to existing ones. Therefore, a sincere assertion, for example, is characterised such that for an utterance where S addresses a declarative sentence with propositional content p, to H in an initial state where S believes p, S's belief persists and H adopts a mutual belief which is not contradictory to one in existence.

BEFORE: $B_S p$, AFTER: $B_S p$

and MB (H,S,p) or $B_H p$, & $B_H B_S p$, & $B_H B_S B_H p$ &.....

A successful lie can be characterised similarly with S's belief again persisting, and H adopting the new one, it not being contradictory to one in existence.

BEFORE: $B_S \sim p$, AFTER: $B_S \sim p$

and MB (H,S,p) or $B_H p$, & $B_H B_S p$, & $B_H B_S B_H p$ &.....

Irony is characterised with S believing $\sim p$ before the utterance and H believing this. Since beliefs persist, both S and H believe the same as they did, even after S's declarative utterance that p. H does not adopt the mutual belief of the earlier characterisations, because it is contradictory to his existing belief.

BEFORE: $B_S \sim p$, AFTER: $B_S \sim p$

and $B_H B_S \sim p$ and $B_H B_S \sim p$

Morgan (1987) criticises this analysis of irony on the basis of its assumption that ironic expressions can never inform. The hearer must already have a belief about the speakers beliefs regarding p. This is not necessarily always the case and makes ironic utterances a "total waste of time" (Morgan, 1987). In fact, Morgan's more general criticism is that Perrault's belief transfer rule is a "great over-simplification" (Morgan, 1987). This rule embodies the component of the persistence theory of belief which states that beliefs can be transferred from one agent to another, as long as they are consistent with existing ones. Perrault acknowledges this point:

> "Ideally, one would like a theory in which it is possible for one agent's beliefs say, to change depending on how strongly he believed something before the utterance, and how much he believes what the speaker says. We cannot give such an account in detail, so we rely on something simpler"

(Perrault, 1987).

One consequence of this rule is that agents cannot use dialogue to "change each others minds". Obviously, a belief transfer rule more in line with Perrault's ideals was crucial to this research.

1.4 Conclusions

The aim of this research was described in the introduction as being to develop a theoretical framework for computational models of cooperative dialogue. The focal issue is the acknowledgement not only of the role of conflict in multi-agent cooperation, but also the importance of dialogue in this. An existing pragmatic theory of dialogue, which considers communicative action to be grounded within a theory of multi-agent interaction, has been taken on and developed according to the objectives of this research. Cohen and Levesque's approach to speech act theory (Cohen & Levesque, 1987b) is that theory of dialogue. In the following two chapters, a new theory of multi-agent interaction is introduced as the context component of this theory. It is the context within which agents can reason about dialogue action and conflict and cooperation.

The relevance of this chapter has therefore been to introduce the linguistic theory to the proposed framework; chapters 2 and 3 comprise the theories of rational agenthood and multi-agent interaction upon which this linguistic theory depends. They are the means by which the theory of dialogue described here has been extended to incorporate negotiation as a means of expressing and resolving differences between cooperative agents.

CHAPTER 2: A theory of rational agenthood

2.1 Introduction

This research concerns the development of a theoretical framework appropriate to the computational modelling of cooperative dialogue. The previous chapter described its linguistic background as Cohen and Levesque's recent work in speech act theory (Cohen & Levesque, 1987b). The proposed framework therefore comprises a pragmatic linguistic theory where speech acts are utterances, performed and interpreted within the context of the agents' mental states. According to the advocates of this approach (Cohen & Levesque, 1987, Perrault, 1987), the basis upon which speech acts are both generated and recognised is not according to a taxonomy of acts with specified preconditions and effects, but according to a theory of rational agenthood and multi-agent interaction. By agents understanding the properties of their own and others mental states, how these interrelate, how they relate to speech action, how they characterise other properties of agents and interaction such as cooperativeness, rationality, and so on, action appropriate to this and the situational context can be both generated and interpreted via a variety of surface linguistic structures.

The focus of this research is the development of such a theory of rational agenthood and multi-agent interaction as the basis for cooperative dialogue. It is a development of previous research in artificial intelligence and cognitive science which has drawn on insights mainly from psychology and philosophy concerning the architecture of agents and multi-agents. What is it to be a rational agent which can interact cooperatively with others?

The contrast with previous work lies in its being a theory which considers cooperative dialogue to include negotiation as a means of potentially securing agreement. Conflict resolution is a positive force in the maintenance and evolution of cooperative systems. Disagreements between cooperative agents are not avoided. Agents can potentially be persuaded to "change their minds". It incorporates a strategic rationality, and makes no assumptions concerning the benevolence of agents towards each other. Agents are considered autonomous. They have control over the flow of information in the multi-agent system, both in terms of what they acquire and what they reveal. There are also therefore, no assumptions

regarding sincerity. Dialogue is a game of strategy, and the outcome is flexible in a way which imposed benevolence would never allow. Thus can multi-agent systems as a whole maintain themselves and evolve appropriately in unpredictable and changing environments.

The proposed theory of multi-agent interaction is built upon a theory of individual rational agenthood. The role of this chapter is to discuss the latter, before chapter 3 extends it to deal with issues specifically relevant to cooperative multi-agency and interaction. Together, chapters 2 and 3 therefore provide a descriptive introduction to the issues relevant to the entire theory. Explanations are given of the properties of agents in terms of their derivations; some from other inferential studies in communication, others from disciplines such as philosophy, game theory, and psychology. Justification is offered by critically evaluating related research in artificial intelligence, in terms of the consequences of their assumptions as to the nature of cooperative, rational, autonomous multi-agent interaction. In chapter 4, details are given of the language with which the properties of agents are then formally expressed, tested, and evaluated as a basis for cooperative dialogue, in the following chapters.

2.2 Rational, intelligent agenthood - the agent model

> "Reasoning about the cognitive state of other agents is an essential part of intelligent behaviour" (Konolige, 1986).

Our own actions in a multi-agent world are based on considerations such as:

> "what others want, fear, know, intend, and so on; and there is every reason to expect, as we develop more sophisticated, autonomous AI systems that interact with humans and each other, that they will also have to reason about at least some of these concepts" (Konolige, 1986).

In other words, multi-agent interaction is assumed to require recognition and understanding of what it is to have mental states, and how actions express and alter these. The folk psychology view which has

generally been adopted in artificial intelligence, is that an agent's mental states are identifiable and accessible structures, representing information about themselves, other agents and the world. Some of these are cognitive in nature and represent the agent's beliefs and knowledge. Others are conative representing desires, wishes and wants. Yet others represent affective issues such as the agent's values, likes, and preferences (Kiss, 1986).

This notion of mental states as identifiable structures is not without contention. In his book about the relationship between folk psychology and cognitive science, Stich refers to several theorists whose work involves cognitive modelling, and who do not propose any physically or functionally isolatable components corresponding to beliefs or desires (Stich, 1983). He quotes Winograd, and Minsky as examples. For example, Minsky's "Society of Minds" view is that none of the components of mental models have meanings in themselves; meaning emerges from "great webs of structure". Therefore, no part can correlate with explicitly represented beliefs (Minsky, 1981). Dennett's work (Dennett, 1978) is also referred to, in which beliefs and desires are argued as being merely instrumentalistic concepts. They therefore need not correspond to any physical or functional state. Predictions of behaviour can be made by ascribing beliefs, desires and intentions, but these need not be actually represented within the agent.

However, in accordance with the majority of work in this area, an important assumption about the nature of agenthood being made here is that agents have cognitive, conative and affective representations of the world. The question to be answered in the next section is how - in what form - are these represented? Following this, the nature or characteristics of each are discussed. What are the properties of beliefs, desires, likes, dislikes or preferences? Finally, section 2.2.7 concerns agent rationality; the relation between mental states and action.

2.2.1 Representing mental states

Mental states are represented here as propositional attitudes. This term was first used by Russell, and expresses a relation between the agent and some proposition. For example, "Ben believes that it rained yesterday". The proposition that it rained yesterday has a truth-condition. It can either be true or

false. It can be true or false in the real world or in some possible world(s)[1].

A propositional approach has been adopted because it is a semantic means of defining mental states. Propositions express meaning. For example:

> "'Jack and Jill have one parent in common' expresses the same proposition as "Jack and Jill are step-siblings'" (Haack, 1978).

Mental states are therefore defined with respect to notions of truth, as opposed to the alternative approach which is to represent them as relations between the agent and a sentence. Such a syntactic approach differentiates beliefs and goals not according to their meaning, but according to the grammatical string of expressions or symbols.

Associated with these alternative theoretical approaches to the representation of mental states, are alternative languages or means of expression. Mental states represented as data structures could be either expressions of a computer language or of a formal language. This research has adopted a formal approach for reasons outlined in full in section 4.2, and the different formal approaches are discussed also in chapter 4. Inevitably some properties of the agent model described here are a direct consequence of the chosen means of representation and expression. The reader will be made aware of these as they arise in the course of this chapter, but full discussions of these issues are to be found in chapter 4. The next few sections concern the nature of the three types of attitude. Reflecting the relative volume of research into the cognitive or epistemic attitudes of knowledge and belief as opposed to the others, these are dealt with first.

2.2.2. Knowledge and belief

In the philosophical literature, beliefs as propositional attitudes are distinguished from beliefs as

[1] details on possible worlds in chapter 4.

psychological states. However, in "commonsense" or "folk" psychology, there are elements of both. People are said to perform actions because they have a certain belief. This implies beliefs as psychological states which play a causal role. However, it is also the case that actions are explained in terms of what is believed, or the content of the belief. (Engel, 1984). Belief can be described as a disposition to act, given certain relations to other states (Engel, 1984).

> "Believing is a disposition that can linger latent and unobserved. It is a disposition to respond in certain ways when the appropriate issue arises.....To believe that frozen foods will thaw on the table is to be disposed, among other things, to leave them on the table only when one wants them thawed" (Quine,1970).

Preferences, for example, are defined in section 4.4.4.1 in terms of a belief which expresses a relationship between certain beliefs and a goal. Preferring to go out than stay in under certain circumstances, is defined as a disposition such that if those circumstances should arise, then a goal is generated to go out.

The precise properties of knowledge and belief as the cognitive components of the agent model used in this research, are derived from their representation using the Hintikka schemata (Hintikka, 1962) as modal operators with a possible-worlds semantics. This particular approach to the representation of the epistemic attitudes is described and discussed in section 4.3.1. The resultant properties include one property which distinguishes knowledge and belief. This is that agents know only propositions which are true, but they can believe propositions which are in reality, false. The other properties apply equally to knowledge and belief. For example, if an agent knows or believes p, she knows or believes all the other propositions which are also true in such a state of the world where p is true. Agents are then also modelled as having unlimited resources with which to draw all the possible inferences from all that they know and believe. They therefore know all the logical consequences of what they know and believe all the logical consequences of what they believe. Agents are introspective; they know what they know and don't know, and believe what they believe and don't believe.

It is obviously the case that these properties need refinement. Firstly, there should be a distinction

between those propositions which agents consciously or explicitly know and believe, and those which are not actively held to be true but merely follow from what is believed. This could then lead to the possibility of reasoning with only some beliefs. In fact, it would seem a desirable property of agenthood that agents reason at different times with different sets of beliefs which should be internally consistent, but need not be consistent with each other, as long as they are accessed at different times. Attempts to alter the possible-worlds approach along these lines (Levesque, 1984, Fagin & Halpern, 1985) are described in section 4.3.1.2. It is also less than feasible that agents know all that they don't know, and have beliefs about what they don't believe.

Another criticism stems from the fact that all knowledge and belief in this model simply exists; it is encoded as a vast set of sentences concerning infinite sets of propositions. Wilks (1983) offers a computational alternative in which complex beliefs about other's beliefs can be constructed when required from "bottom level" structures. He gives as an example a belief about what Reagan thinks Begin thinks of Gaddafi. In possible worlds models, such a belief would have to be already in existence and stored.

A particularly important property of belief, of relevance to issues of belief revision discussed in the next section, is that beliefs can be held with different strengths or intensities. Appelt refers to beliefs being held with varying "degrees of certainty" which occur in response to the acquisition of new information (Appelt, 1985). Quine refers to beliefs like the charge of a battery; they "may last long or briefly". Some beliefs are retained for life and others may be abandoned easily in the face of adverse evidence (Quine, 1970). The work of Rokeach offers a psychological explanation for this in terms of centrality of beliefs. Fundamental beliefs are more central to the agent in having many connections with other beliefs. Those on the periphery have fewer connections and are therefore less resistant to change (Rokeach, 1975). Therefore a stronger belief can be envisaged as connected with not only more other beliefs, but beliefs also more central to the agent. In that way, a belief which implies two other beliefs, as long as the two are strongly held convictions, will be maintained in preference over one which relates to several less tenacious ones. For example, a belief that my sister has not committed a crime will be retained in the face of several beliefs relating to circumstancial evidence that she has, if my sister's honesty is a central belief.

Both Appelt and Perrault refer to the difficulties of formally accounting for varying strengths of

belief. Appelt's solution is to avoid beliefs altogether and "consider the simplest cases first". He uses knowledge as the only cognitive attitude (Appelt, 1985). Perrault offers his persistence theory of belief referred to in sections 1.3.5 and 2.2.2.1, as a way to initially "rely on something simpler" (Perrault, 1987). Cohen and Levesque do not discuss this problem; beliefs all at one level of intensity are assumed (Cohen and Levesque, 1987a, 1987b). As mentioned earlier, this work has adopted Cohen and Levesque's adaptation of Hintikka's representation of belief in the formal model of agents, and therefore here too, beliefs are represented at only one level of intensity. However, the notion of varied strengths of belief is very important to the psychological basis for preferences. Along with the notion of goals as also being held at varying strengths, this idea is included in the theory of preference, which is explained in section 2.2.3.

2.2.2.1 Belief revision

In the previous chapter, speech actions were said to convey information about the speaker's mental states; in particular the intention that certain changes be effected to the hearer's mental states. Such changes to an agent's belief states may involve simply addition to the beliefs already there. On the other hand, the information recognised by the hearer and concerning intended changes may in some way contradict an existing belief.

> "Though many of our beliefs are here to stay, at other points the body of our beliefs is perpetually in flux. Primarily this is because our senses keep adding information. This simple addition of information, at the sensory end, issues in change in the body of beliefs that is not to be equated with simple addition of beliefs. For one thing, beliefs get crowded out and are simply forgotten. This happens promptly to the host of trivialities such as the chirp of the bird and the chug of the motor. For another thing, more to the point of our study, beliefs still vigorously present and not to be forgotten can come into conflict with the new arrivals and be forced from the field.....We can no longer believe all of a set of sentences to be true once we know them to be in contradiction

with one another, since contradiction requires one or other of them to be false." (Quine, 1970).

The idea of beliefs getting "crowded out" and forgotten may or may not be relevant to computational applications not pertaining to be psychologically plausible models. Machines can have "idealised" memories with immediate and direct access to all beliefs, or they may only keep the concepts constantly in use directly accessible, and have the rest in secondary forms of storage. However, the notion of beliefs in conflict and therefore requiring resolution is appropriate. In the current model, agents cannot believe p and not p simultaneously. What should be the basis upon which either a new communicated belief be adopted, or old beliefs maintained, in such circumstances?

As mentioned in the previous chapter, Perrault's answer to this question is embodied in a persistence theory of belief. This states that existing beliefs persist over time, and new beliefs are only adopted if consistent with those already in existence (Perrault, 1987). Thus beliefs are never actually revised; recognised, inconsistent beliefs are simply not taken on. Perrault acknowledges that this is a simplification adopted for specific research purposes. It is a simplification which is unsuited to the objectives of this research programme. Conflicts can only be truly resolved using dialogue, if the dialogue which presents new information can actually effect changes to mental states, not merely add to those in existence.

With Cohen and Levesque's framework, an agent can apparently stop believing p for example, when some event occurs, the result of which is that p is no longer true in all the possible worlds compatible with her beliefs. Agents believe all the logical consequences of their beliefs and therefore, if believing q implies not p, and q is believed following some event, then p will no longer be true in all the possible worlds compatible with what the agent believes (Cohen & Levesque, 1987a, 1987b). However, what is the basis for q being adopted, and thus p rejected? Can agents adopt q because someone tells them q? Is recognition of another agent's intention that q be adopted, sufficient? In terms of goal adoption, recognition of another's goal is only sufficient, if this goal "does not conflict with one of his own" (Cohen & Levesque, 1987b). It is assumed that Cohen and Levesque consider belief adoption in dialogue to also conform to such a basis. Another's goal that one believe p is adopted only if one currently does not have the goal to believe not p.

The research described by this thesis, requires that the conditions under which an agent will acquire mental states during multi-agent dialogue interaction reflect the agent's autonomy over the matter. This is especially important to the use of dialogue as a means of resolving conflicts. Agents can choose whether to adopt communicated beliefs or not, and the basis for this is not consistency with existing beliefs. Beliefs and goals are acquired during dialogue conditionally upon either evidence, or preference.[1]

Evidence of truth in the world is an indisputable basis for belief revision. If the hearer believes that the speaker wants her to believe p, and believes also that the speaker knows p, then she believes this because she has evidence of p being true in the world. An example would be a child crying that she is in pain. She wants her mother to believe that she is hurt. Perhaps the child frequently "cries wolf". However, if the utterance is accompanied by the evidence of a grazed knee and blood, the mother will believe her daughter knows she is in pain; it is true in the real world. She will also then adopt the intended belief. Frequently however, beliefs are adopted and/or held without any evidence. Sometimes evidence is gathered retrospectively in order to justify and substantiate changes in beliefs already made. If this is the case, then what was the original basis for the belief adoption? The answer offered here to that question, is a preference.

> "Would it still be taken to support the belief in question if we stripped away all motives for wanting the belief to be true?" (Quine, 1970).

Preferences are discussed in the following section.

2.2.3 Preference.

Preferences are attitudes which reflect a certain affective aspect of agenthood. They are attitudes

[1]The reader is referred to chapter 6 for the detailed account of this theory of belief and goal adoption in dialogue between autonomous agents.

which describe a relation between an agent and a pair of propositions; agent x prefers p to q, for example. They have been defined in chapter 4 in terms of the other primitive concepts of belief and goal, much as Cohen and Levesque use beliefs, goals and action to build a definition of intention as a molecular concept.

A preference describes the relationship between a belief about a pair of propositions, and a goal. Preferences are indicators of which out of two alternatives to choose, should the appropriate circumstances occur. For example, if I have a preference to eat strawberries with ice cream than eat them without, this will determine my goal, and correspondingly my actions, should the circumstances arise wherein I have such a choice. Preferences are formally defined in section 4.4.4.1. They specify which proposition to retain and which to reject, should a particular situation arise. Agents are assumed to have the machinery to be able to compute these. This machinery derives from the psychological background to preferences which is suggested as relating to both the numbers and the strengths of supporting beliefs and goals, for each of the options. Firstly, a proposition will be preferred to a particular contradictory one, if it satisfies more goals and is consistent with more beliefs than the alternative. For example, the belief that this coming Sunday will be spent working may be preferred to a belief that it will have been spent going to the park with the children, because it satisfies several goals at once. Examples of these may be getting an overdue paper written whilst making everyone notice how hard-working you are, and not letting down a colleague. On the other hand, if having gone to the park only satisfies one goal, such as pleasing the children, but this goal is much more fundamental and stronger than the other goals mentioned as was explained for beliefs in section 2.2.2, then the preferred belief is that next Sunday will be spent in the park rather than working. The preference dictates which goal be generated, when actually faced with the choice.

It is important that agents believe themselves and others to understand this basis for preference, and hence also belief and goal adoption or rejection in dialogue. This information can then be used in attempts via dialogue to induce another to adopt or drop a particular belief or goal. In the example above, if a colleague about to be abandoned to work alone on Sunday believed that promotion was yet a stronger goal of the other agent than pleasing his children, by performing an utterance inducing a belief of the possible detrimental effects on promotion possibilities, his colleague's preferences would be

operating in an altered context, which might therefore result in a different action.

This basis for preferences provides a "moral" element for this framework of dialogue in which speech actions can be generated strategically, and with no assurances regarding issues such as sincerity. Imagine for example, a situation where x believes that she could either assist her friend who is requesting such help, or satisfy a contradictory goal of her own. The friend knows that the basis for x's decision is her preference between these two options, and the basis for this is the number and strengths of other goals and beliefs which potentially would be satisfied in each case. If she believes the desirability of altruism towards a friend is a strongly held belief of x, then she believes x's preferences will dictate that the reply to her request will be made in her favour, and importantly this occurs without imposed benevolence. x's assistance is an autonomous and rational decision. This issue arises again in discussions regarding the nature of cooperation in chapters 3 and 5.

2.2.4 Wants and goals

The conative component of this model of agents is the propositional attitude, goal. Agents' goals are modelled similarly to their beliefs, using possible-world semantics. The properties of goals are therefore also determined by this particular formal approach, a more detailed account of which can be found in Chapter 4.

Goals characterise what is implicit in the agent's desires. Having a goal that p, describes what the world would be like if p were true. This means that implicit in an agent's goals are all the logical consequences of those goals, just as having one belief means all its logical implications are also believed. A goal is a state of the world, and it is a state of the world the attainment of which, is desired by the agent. Attainment of the state thus satisfies the goal.

The term "goal" is used frequently in AI. Planning systems plan sequences of actions to achieve desired or goal states. The properties of goals - what it means to have a goal, are rarely discussed. Goals are broken down into achievable sub-goals. If a goal exists and is achievable, it is to be satisfied. If a goal cannot be satisfied by one means, then backtrack and try another. With multiple goals, more

important goals can be prioritised according to a "weighting" algorithm, and the more goals satisfied, the better. However, if the aim is to characterise the conative element in a model of agents, using goals in this way gives an impoverished picture.

Firstly, agents have multiple goals and their relation to these desired states of the world are not necessarily all of the same type. Some states may be desired, but the agent believe them impossible to attain. For example, having a desire to buy a Mercedes, but this being an unaffordable objective. On the other hand, it might be believed possible, as well as desired but not an objective to which the agent is committed in the sense of actually making plans to satisfy the goal. Commitment is a very important notion. There needs to be a distinction between goals which the agent will plan to satisfy, and others to which she may be currently less committed for some reason. Therefore, although rational action can be defined as action performed by an agent as a means of achieving goals, it is also an expression of the rational agent's commitment to that particular goal. Cohen and Levesque define a committed, or persistent goal as one which the agent will only give up when achieved, or if the agent believes it impossible to achieve, or if the reason for the goal is no longer true (Cohen & Levesque, 1987a, 1987b). These ideas have been adopted here.

In fact, the existence of a goal already indicates some level of commitment on the part of the agent. From the possible-worlds model, agents goals are consistent; an agent cannot have a goal for p and a goal for not p at the same time. This means that goals do not represent desires which may be conflicting yet simultaneously held, such as wanting one's cake and eating it too. The use of goals is in fact a means of avoiding this problem. A more realistic or psychologically plausible candidate for the primary conative attitude would perhaps be a wish or a want (Kiss, 1973). However, in accordance with the others who have followed this path of formal reasoning such as Appelt (1985), Cohen and Levesque (1987a, 1987b), and Perrault (1987), agents reason with a set of consistent desires.

> "We assume that once an agent has sorted out his possibly inconsistent desires in deciding what he wishes to achieve, the worlds he will be striving for are consistent" (Cohen and Levesque, 1987a).

In accordance with the adoption of Cohen and Levesque's approach, this model restricts goals to only

those states of the world which are a subset of the worlds believed possible. This is justified as a "realism constraint". If a rational agent believes he will be dead in two months time, he will not buy a plane ticket on the basis of a goal to be in Miami, three months hence (Cohen and Levesque, 1987a). This points again to wishes or wants possibly being a more appropriate primary conative attitude than goal (Kiss, 1973). In this case then, a desire to go to Miami would still be feasible, if believed impossible. In fact, this whole discussion concerning different types of conative attitudes further indicates the desirability of being able to distinguish attitudes on a basis of their relative strength or centrality to the agent, as was discussed in relation to beliefs in section 2.2.2.

Another important question, is where do goals come from? As with beliefs, agents are modelled as having an existing set of goals. Agents can also adopt others' goals as well as beliefs during dialogue, according to the conditions mentioned in section 2.2.2.1 and described in detail in chapter 6. Agents can generate sub-goals from existing goals. There is little existing work however, which acknowledges the role of goal generation, other than as sub-goals to existing, pre-programmed goals. Wilensky is one exception. He has a Goal Detector comprising a Noticer component to "notice" environmental change (Wil ensky, 1984). Sloman suggests a "motivational store" accessible to the system as a basis for goal generation amongst other things, yet no practical details are given (Sloman, 1978). This model offers some advantage in this respect. Preferences are described in sections 2.2.3 and 4.4.4.1 as prescriptions for goal adoption or rejection, in the event of certain alternative conditions being true. If the agent believes that one of two options may become true in future and has a preference for one of these over the other, then a goal is generated for this to be the case. This framework has concentrated on demonstrating the relevance of preference in relation to belief and goal adoption following recognition of another's communicative goal in dialogue, but this principle is also generalisable to other contexts.

Finally, the term communicative goal needs to be clarified. It was said earlier that rational action is purposive in being consistent with the agent's goals. Following from Austin (1962), rational speech action is a special case of action. Therefore, speech action generated in order to cause effects on agents mental states is also purposive in being consistent with the speaking agent's goal or desire to induce such effects. In this theory the term communicative "goal" is used when referring to the desire to induce changes in the mental states of others. The speaker also has a goal that the receiving agent recognise

this goal. This is merely a terminological variation from the Gricean component of the speech act theory as explained in Chapter 1. Grice's term is "intention", yet this term has a different interpretation here, as explained in section 2.2.6.

2.2.5 Interests

Interests[1] have been incorporated into this model to represent a type of goal which is believed by the agent will be achieved. Its formal definition and further explanation can be found in section 4.4.4.2. It is an additional conative state which is particularly appropriate to situations of multi-agent conflict. In such contexts, believing that one's plan will actually be successful and not merely feasible, may be an important component to the generation of the plan. An example is a situation where one agent wants another agent to believe what is in fact not the truth, when their discovery of this would have dire consequences. The speaker must assess carefully that the plan will work, before telling the lie; it is not enough that it might work. This still however, does not mean that in reality it will work; the agent only has to believe this to be the case.

It should be emphasised that assessing whether a desired state is actually likely to be attained before planning to achieve it, as opposed to just not being impossible, can be costly in terms of processing resources. In many dialogue contexts, it is not appropriate. For example, x's plan is to go out somewhere tomorrow. She has a goal for y to ask x to come over tomorrow. As long as she doesn't actually believe it impossible that y will do this, then she can make a commitment to achieving that goal, and if it fails, then she can simply try something else, such as asking z to take her out. When a goal becomes believed to be impossible, then according to the conditions described by Cohen and Levesque for commitment to a goal, the goal is dropped. The relevance of the agent reasoning about action in relation to her interests is specific to strategic multi-agent interaction. Strategically rational

[1]The term "interests" has been adopted independently of its standard useage in other disciplines, such as moral philosophy or game theory.

action requires reasoning with not only one's goals, but the likelihood of success - and from this, one's expectations of the other agent's subsequent response. This is because the desired state of the world relates to the other agent's mental states, and the result is not as predictable as in the single agent's dealings with the physical world. In conflict situations, the consequences may be crucially important; if the plan fails, there may be no going back and trying again. Game theory, for example, acknowledges this in the form of probabilities being incorporated into the determination of preferred outcomes before selecting moves in games involving conflict between multiple players. In this framework, purposive action is strategically rational when the acting agent believes the goal will be achieved. This issue being more appropriate to multi-agent interaction, it will be elaborated in Chapter 3.

2.2.6 Intention

Agents may have goals to induce changes in the world, such as effecting changes to others' mental states. However, their intentions are intentions to act. Thus, whereas knowledge, belief and goals are types of propositional attitude whose content is a proposition, intention differs in that its content is an action. It also differs in its properties relating to relationships with other attitudes, as well as time and action, and therefore it cannot be analysed in isolation.

The view of intention in action derives from a rich philosophical background in action theory and practical reasoning (Castaneda, 1975, Davidson, 1963, Anscombe, 1963). It is an element in Cohen and Levesque's theory of agenthood and rational action which they demonstrate as the basis for communication (Cohen & Levesque, 1987b). They claim that rational behaviour should not be analyzed in terms of beliefs and desires alone. Intention is also an important component, which although related to them, is not reducible to them (Cohen and Levesque, 1987a). Philosophical studies of intention have frequently tried to reduce intention to a combination of these two primary attitudes however. Brand refers to this as the reductive analysis or (DB) which can be summarised as follows:

S intends to do A at t iff S desires to A at t and S believes that she will A at t (Brand, 1984).

Bratman, whose work provides the philosophical foundation for Cohen and Levesque's formal theory of intention (Bratman, 1987), claims such definitions are insufficient. Rational behaviour is analyzable not only in terms of beliefs, and goals, but involves another mental state which incorporates a notion of commitment - this being intention. His view is that intentions are parts of plans; they are plans for future action. His reasons are as follows: Firstly, agents cannot continually waste resources weighing up competing desires from which to generate intentions. Some of those desires will be held with greater commitment. Brand uses similar arguments in suggesting that an improved definition to the above would be for example, that an intention to perform A at time t requires that there be no B such that the agent's desire to do B at t is stronger than the desire to do A (Brand, 1984). Secondly, coordination of several future actions requires a notion of commitment to those intended.

Intention is modelled here in accordance with the definition provided by the work of Cohen and Levesque, as a type of persistent goal. Intention is a commitment to believing one is about to do an action and then doing it. Thus, from commitment to a goal such as wanting to induce a belief p in another agent y, an agent x may generate an intention to perform a particular speech action. The goal of the speech action is that y recognise x's goal for y to believe p, which is a subgoal to the original goal. x's intention is a commitment not to any goal, but to believing she is about to do that action and doing it. Rational actions can be viewed therefore as those which not only satisfy goals, but the attainment of intentions. Rational agents adopt only those intentions they believe to be achievable.

2.2.7 Rational action

Agents with the properties outlined in the previous sections can generate actions. With knowledge of these properties, there can be expectations of an agent's actions, if the principles whereby they relate to action are outlined. Thus a theory of rationality is a predictive theory. It provides a prescription for action; it provides a basis upon which agents' actions can be understood.

In order to describe the properties of agenthood, reference has already been made to their relations to

action: Rational action is purposive, or goal-directed; rational action is not only consistent with the agent's desires, but consistent with her beliefs, such as a belief that the desired goal state is achievable; rational action is consistent with intentions; and rational action expresses commitment.

There are however, different analyses of rationality (Harrison, 1979). The above indicates that this research has taken an evaluative approach, where actions are selected as rational in contrast to others, dependent on consistency relationships with the individual's other existing mental states. Another evaluative approach would be to consider particular beliefs or desires as rational, and then evaluate behaviour as rational or irrational in relation to these. This requires a view of rational action as equivalent to what is "normal" or "natural". Finally, a descriptive approach considers all behaviour rational; it stipulates humanity. This "holism" of the mental is described by D. Davidson, where complete sets of beliefs and desires are attributed on the assumption that they inevitably fit together rationally.

Allwood describes rationality as "primarily an instrumental concept. It designates a manner of thinking or acting to reach a certain goal." (Allwood, 1976). His approach is evaluative, and relates rational action to a consideration of consistency with the individual agent's goals and beliefs, but with reference to some kind of independent assessment of those beliefs or goals. He refers to the "normal" rational agent, as well as saying:

> "We are not making the claim that agents act rationally, but only the weaker claim that agents act in a way that seems rational to themselves" (Allwood, 1976).

Allwood claims an individual's rationality is based upon his own motives and presuppositions about the world. He summarises normal rational agenthood in seven principles which agents assume they and others adhere to. The first is the basis for all the others. It says that agents are assumed to be normal and rational. The following two principles elaborate what this means for the actions of agents; actions are assumed to be intentional and purposeful, as well as voluntary. The next two principles elaborate "normality" assumptions; normal agents act with motives, such as needs or desires, and normal agents do not act so as to decrease their pleasure or increase their pain. The final two principles explicate the

rational component; the actions of a rational agent are selected on the basis of being the most adequate and efficient ways of achieving the purpose for which they were intended.

> "Given a certain goal, the most rational way to reach it is the way with the least costs involved" (Allwood, 1976).

In "A Treatise of Human Nature", Hume also refers to rational action as "the most efficient means to a desired end." (Hume, 1978). Finally, the principle of competence states that rational actions are those which the agent believes will possibly achieve their intended purpose.

Allwood's work is unusual in being one of very few attempts to explicitly summarise the traits of a rational agent. His motivation, similarly to that of this research, was in relation to development of a theory of communication from an understanding of the nature of multi-agent interaction (Allwood, 1976). Artificial intelligence systems which plan actions, linguistic or non-linguistic, embody principles of rationality. However, generally these are attributed without consideration of this as a separate issue. It is merely in the design of systems with certain properties, such as knowledge of which actions can be performed, actions being performed in order to satisfy goals, the more efficient means of goal satisfaction being the one selected, and so on, that rationality is conferred upon the system; its principles are therefore simply embedded within the system architecture.

Cohen and Levesque's work aims to provide a formal specification of the properties of rational agents for artificial intelligence applications which concern agents reasoning about each other, as for example when communicating (Cohen and Levesque, 1987a, 1987b). Their principles of rationality for the single agent, can be summarised as follows: Agents' actions are purposive in being consistent with their goals, and goals are consistent with the agent's beliefs. Agents choose the possible worlds they would like to be in and they only choose worlds they believe to be attainable. Agents choose amongst their goals what they believe to be inevitable. Agents do not try forever to achieve their goals. If a goal is believed to be impossible to achieve, it will be dropped. Likewise, goals are not deferred forever. The conditions for a goal to be dropped are that it be achieved, believed impossible to achieve, or the reason for the goal is no longer true. The opposite of these three conditions represents commitment to a goal.

Agents' actions are consistent also with their intentions. These are commitments to do an action or to achieve a state of affairs.

This framework also incorporates these ideas on rationality of Cohen and Levesque's, but with two important additions. Firstly, there is an extra condition under which a goal can be both adopted and dropped - preferences. Acting to satisfy a goal which is inconsistent with one previously held is rational, if the new goal is generated from the existence of a preference, in conjunction with certain pragmatic circumstances. The preference expresses a relationship between beliefs and goals whereby such a goal would be generated when faced with those particular circumstances. In addition agents can reason about generating purposive communicative action in relation to goals which the agent believes will be achieved. These are their interests.; goals are merely believed achievable, or not impossible. This second point is of particular relevance in strategic interaction where it may be appropriate to believe that a particular goal will be satisfied before planning to satisfy it. In some conversations however, as long as an objective is feasible, such extra reasoning is a waste of computational resources.

Communication is obviously an activity involving more than one agent, and there are principles of rationality which apply specifically to interaction between multi-agents. The basic principles of rational action which have been discussed here, have to be extended to cater for the multi-agent context. The following chapter is devoted to the issues relevant to multi-agent interaction which particularly relate to cooperative dialogue. The role of interests in rational interaction for example, is discussed there in more detail. Before moving on to discuss the multi-agent case however, it needs to be pointed out that the author is aware that it is an idealised notion of rationality which has been presented here for the single agent, and which is consequently embedded in the theory of multi-agent interaction. Cherniak formulates this as:

> "If A has a particular belief-desire set, A would undertake *all* and only actions that are apparently appropriate" (Cherniak, 1986).

The agent therefore has a perfect capacity to choose actions appropriate for his belief-desire set, which in turn requires him to make any deductive inferences from his beliefs. Cherniak suggests that although

this is a convenient simplification, it is:

> "unacceptably stringent in important ways; for in fact this rationality condition is generally unrealisable" (Cherniak, 1986).

The reason is that there are limits on cognitive resources, and he claims that these are not just human limitations.

> "They would be just as unavoidable, for example, for a creature that had available the resources of the entire galaxy until heat-death of the universe" (Cherniak, 1986).

He offers instead, a minimal rationality condition :

> "If A has a particular belief-desire set, A would undertake some, but not necessarily all, of those actions that are apparently appropriate".

There is also a minimal inference condition, minimal consistency condition and minimal deductive ability .

> "Not making the vast majority of some feasible inferences is not irrational; it is rational" (Cherniak, 1986).

This is justified if deducing all the consequences of a belief, some of which are trivial, prevents other inferences which may be crucial to survival. Simon similarly discusses the value of "satisficing" as opposed to "maximising" in decision theory (Simon, 1957). Pollack, Israel and Bratman (1987) refer to the significance of these points made by both Cherniak and Simon, and offer a proposed architecture for rational behaviour in resource-bounded agents. These are agents which are "unable to perform arbitrarily large computations in constant time" (Pollack, Israel and Bratman, 1987). Their architecture includes a

filtering process to constrain the overall amount of necessary practical reasoning, based upon a theory of the functional role of plans in this.

It seems inevitable that future practical systems modelling aspects of agenthood will need a rationality which acknowledges cognitive limitations, and restricts deductive inference to that which is minimally necessary. An example of an attempt at this is the work of Ramsey (1987). He offers a set of inference rules for reasoning with other's knowledge, which are not complete, but quick and effective. They mimic the reasoning which might be performed by another person, as opposed to analysing how facts are constrained by others (Ramsey, 1987). Konolige's deduction model (Konolige, 1986) employs deductive mechanisms for deriving some but not all logical consequences from a core set of beliefs (Konolige, 1986b). However, this research has adopted a possible worlds approach to the modelling of agenthood for the reasons given in section 4.2. In particular, in order to develop Cohen and Levesque's (1987a, 1987b) work and in so doing, concentrate on aspects of multi-agenthood and interaction such as cooperation and control of information flow, as essential elements of a framework for cooperative dialogue. The rationality idealisation is acknowledged, but therefore currently remains.

CHAPTER 3: A theory of cooperative, multi-agent interaction

3.1 Introduction

The description of rational agenthood given in the previous chapter, is extended here to encompass those properties appropriate to interacting multi-agents. Because of the stated research interests , the emphasis is on those properties which are relevant to cooperative interaction by means of verbal communication. The nature of multi-agenthood, what it is for multi-agents to be cooperative, the role of conflict in cooperation, and the role of dialogue in conflict resolution and cooperation - these are the issues which are addressed in this chapter. This is an introductory and descriptive account however. It includes justifications and comparisons with previous work, and presents a complete story. Those aspects which comprise the focal issues for this research are then developed and elaborated in chapters 5 and 6.

3.2 Rational, intelligent agents in interaction - the nature of multi-agent systems.

A multi-agent system comprises multiples of agents in a common environment. A system is defined in system theory, as a phenomenon consisting of components in relationships to one another. Each agent and the environment are therefore interrelated; no element can operate independently as an individual unit. The structure of the system is defined according to its components. According to Kiss, these components are objects, with both temporal and spatial relations holding between them (Kiss, 1987). If the process-like aspects of systems are emphasised, the relations of interest are temporal ones. Kiss is currently investigating the notion of structure as analysable further, in terms of constraints on possibilities.

Doran (1987a) offers a process-oriented definition of multi-agent systems. He defines a process as an entity whose structure varies in time. Process systems are "a collection of processes which interact and influence one another in a limited way". (Doran, 1987a). He then defines multi-agent systems as a collection of actors or agents, interacting in a common environment. They are specialisations of process

systems, each of which is itself a process system, and which may also be a component of a larger structure. The environment itself is also a process within the system (Doran, 1987a). A more detailed definition of a process is offered by Kiss, as a series of events and states occuring within an occasion. An occasion is defined as an arbitrary delimited spatial- temporal zone. An action is a class of events which results from the activity of agent(s) (Georgeff, 1984). Events, or changes in the world, are caused by those actions (Kiss, 1987). In other words, it is the actions of agents which cause changes to the states of processes, these being the multi-agent system, the agent or the environment. In a multi-agent system, where each agent is not only a process in itself, but a component of the larger process, the actions of agents result in events or changes to the mental states of the component agents - themselves and others, as well as physical aspects of the environment.

From the description of rational, intelligent agenthood given in the previous chapter, the causes of action are the mental states of the acting agent. In a multi-agent system, these mental states must relate to the whole system, which comprises not only the acting agent itself, but also other agents, and the environment. For example, if an agent has a goal to induce a particular belief state in another agent, that agent being rational must have a belief that the other agent does not already possess the desired belief state. Agents have attitudes such as beliefs and goals, which relate to the beliefs and goals of other agents. (BEL x (BEL y p)) for example, says that x believes that y believes p. These being intimately related to agent action, are the causes and effects of change in the multi-agent system.

> "If we wish to understand social action, we must try to elucidate the complex and sometimes paradoxical mental states that can result when two minds simultaneously seek to form representations of each other....Researchers in philosophy and artificial intelligence have shown that everyday communication (eg., in conversation) gives rise to interlocking mental states of great complexity" (Power, 1984).

Examples of mental states which are unique to the multi-agent situation and therefore do not exist in single agents, are mutual or common knowledge, and mutual belief. By definition, these are attitudes which more than one agent must share. A brief description of these and their role in theories of communication, others as well as that presented in this thesis, is given in the next section.

3.2.1 Mutual beliefs and common knowledge.

The role of mutual belief in this framework of dialogue is as follows: Firstly, along with Cohen and Levesque's and other inferential theories of communication, this framework relies upon each agent in a multi-agent system mutually believing that they and others are rational agents. With also some representation of the nature of rational agenthood, and of multi-agent interaction, inferences can then be made with regards to their own action and the actions of others. Secondly, the properties of rational agents have so far been characterised in terms of mental states, such as beliefs and goals. If social concepts or properties relating to more than one agent, are also characterised in these terms, they will at some point encompass mental states specific to the social context or multi-agent system, such as mutual beliefs. Cooperation, conflict and indifference are examples of such social concepts. Detailed descriptions of their precise definitions can be found in chapter 5.

Previous work which attempts to provide formal explanations of such social concepts includes Power's characterisation of cooperation, or "collaboration to achieve a common goal" using another mental state unique to the multi-agent situation, mutual intention (Power, 1984). In his conclusions he cites Bach & Harnish (1979) as also explicitly formulating precise descriptions of social situations in terms of the mental states of the participants, and Lewis (1969), Grice (1969) , Schiffer (1972) and Cohen (1978) as doing so implicitly. Cohen and Levesque's recent work offers cooperative single agents explicitly defined as helpful (Cohen and Levesque, 1987b).

Bach and Harnish use the following definition of mutual belief with which to define social concepts:

It is mutually believed in a collectivity, G that p to the degree to which the members of G believe:

i. that p

ii. that the members of G believe that p

iii. that the members of G believe that the members of G believe that p (Bach & Harnish, 1979).

A social norm is then defined as:

A kind of behaviour, A (in C, which is the kind of recurrent situation in which the norm applies) is a social norm in G to the degree to which

i. the members of G do A in C

ii. it is mutually believed in G that i., and

iii. it is mutually believed in G that the members of G should do A in C (Bach & Harnish, 1979).

Social norms along with practices, rules and regulations, form a "shared conceptual scheme" (Bach & Harnish, 1979). These are the means by which society, described as "a system regulating and organising people's behaviour" (Bach & Harnish, 1979), is internalised in people's beliefs and other attitudes.

There has been a great deal of discussion in recent philosophical and linguistic literature, as well as within artificial intelligence, as to the role of mutual belief and mutual knowledge in pragmatic theories of communication. This follows from the assumption that what an agent believes another agent in a conversation believes, is crucial to the correct interpretation of utterances. Lewis (1969) and Schiffer (1972) first identified the mutual knowledge that it is raining for example (represented as the proposition p) between x and y whilst they stand together watching the rain as: x knows p, y knows p, x knows y knows p, y knows x knows p, x knows y knows x knows p, y knows.....and so on. This definition has been the source of much contention, largely due to the infinite regress of which it is comprised. Sperber and Wilson, for example, concentrate on the interpretation of utterances according to maximal relevance, rejecting all notions of mutual attitudes (Sperber & Wilson, 1987). Amongst various objections they also argue that for an addressee, A and speaker S, to have mutual knowledge that p, they must both have "knowledge of an infinite set of propositions." A has to be able to compute this infinite set of propositions in a finite amount of time (Sperber & Wilson, 1982).

As opposed to entirely banishing mutual beliefs, other researchers have retained the concept, yet offered various solutions to the infinite regress problem. Bach and Harnish for example, simply suggest restricting the levels of embedding to three. Alternatively, there need be no such rigid cutting off point, but the embedding is halted in practice according to the inferential capabilities of the agents (Grice, 1969). Cohen represents mutual belief with a recursive formula which implies the infinite series, but without this being necessarily explicitly formulated (Cohen, 1978). Clark and Carlson suggest the representation of mutual beliefs as mental primitives; they offer a finite mutual belief induction schema (Clark & Carlson, 1982).

A practical solution by Appelt involves the construction of a hypothetical agent which knows

universally known facts - facts which "any fool knows". A hypothetical agent plays the role of "any fool", constructed by a Kernel function. This represents the kernel of knowledge shared by two agents A and B , which is the facts mutually known by them. In this way mutual knowledge between two agents is simply represented by two axioms - one which states that for any two agents A and B, the Kernel(A,B) is equivalent to Kernel(B,A), and the other which says that the set of possible worlds consistent with A is a subset of the possible worlds consistent with the kernel of A and any other agent (Appelt, 1985).

Halpern and Moses offer a solution to common knowledge and infinite regression which involves the introduction of two separate modal operators: common knowledge or c, and knowledge which everyone knows or e. c is the fixed point of e. Evaluating particular instances of the consequences of having common knowledge generates finite approximations of its infinite potential (Halpern & Moses,1984). This idea is equivalent to a potentially infinite recursive function in a computer language, returning a finite value when evaluated with a particular argument.

The nature of the representation of mutual belief used in this research in defining the properties of conflict and cooperation, is according to Cohen and Levesque's definition (Cohen & Levesque, 1987b). This is defined in section 4.4.3.2, in terms of an auxiliary concept, alternating belief. A mutual belief between agents x and y that p, is a regression to the nth level of: x believes p, x believes y believes p, x believes y believes x believes p,.....and so on.

3.2.2 Propositional postures - conflict, cooperation and indifference

The model of agents adopted for this research and described in Chapters 2 and 4, is expressed in terms of the mental states of individual agents, the relations between these and to action. Mutual belief is incorporated as relevant specifically to the multi-agent situation, and expresses a relation between more than one agent and a proposition. It is proposed here to use these properties of agents to characterise the social concepts, conflict, cooperation and indifference. These are collectively termed propositional postures[1]. As with mutual belief, postures are defined as different types of relation between more than

[1]acknowledgements to George Kiss for the suggestion of the term "posture"

one agent and a proposition. For example, (CONFLICT x y p) describes a conflict relation between agents x and y, with respect to the proposition p. It represents a pattern of mental states reflecting x's attitudes to p and to y's attitudes to p, or a mutual belief about this. Importantly however, postures include a conative element - a goal, which relates in some way to the believed attitude of the other agent. In conflict there is a committed goal to change the others believed attitude to p; in cooperation the committed goal is to adopt an attitude, relative to the other's believed attitude to p. Indifference is characterised by a lack of goals with respect to another's believed attitude to p.

An agent with a belief, goal or preference with respect to p, and also a belief or mutual belief regarding some other agent and their relation to p, therefore has a posture with respect to the other agent and the proposition. The nature of the component attitudes, and most importantly the agent's desires or lack of desires with respect to these, determines whether the posture is one of conflict, cooperation, or indifference. The other agent concerned may or may not also have a posture with respect to the first agent and that proposition, and it need not be of the same type. A detailed discussion of all these issues, including the definitions and formal representations of the three postures, is to be found in Chapter 5. The theoretical background to the definition of cooperation follows in sections 3.2.2.1 and 3.2.2.2.

It is claimed that **in order to effectively model cooperative dialogue for real-world applications, multi-agents need explicit representations of postures, characterised in terms of patterns of mental states.** The reason is that if agents can recognise a pattern of mental states that exists between them, and have a representation of an alternative pattern they wish to achieve, then this is achievable if they also have an understanding of dialogue action as a means of effecting those changes. An example could be an agent x recognising a conflict with respect to y and the proposition p, perhaps as a result of some unpredicted change in the environment. x also has a goal for y's cooperation with respect to himself and p. x can potentially achieve a known pattern of mental states, by acting to effect changes to either her or y's mental states, using her understanding of how dialogue can achieve these changes, the conflict which exists and the cooperation desired.

The role of dialogue to the potential alteration of posture will be discussed later in this chapter. For the moment, it is important to recognise the role that explicit representations of conflict, cooperation and indifference have to play, in cooperative multi-agent interaction. To achieve this, more needs to be

said about the nature of cooperation - what it is and also what it is not. There is a lack of recognition in existing work, that conflict is an important component of cooperation. This being the case, there has been no discussion or representation of it in previous cooperative systems. In addition, and because of this, the existing notions of cooperation implicitly or explicitly incorporated in current artificial intelligence research on cooperative multi-agent systems, are simply inadequate to real-life application. In the following section, these existing notions of cooperation will be described and criticised.

3.2.2.1. Existing notions of cooperation in artificial intelligence

Cooperation is an issue in any area of artificial intelligence research which involves multi-agent planning or joint problem-solving, whether concerning human-computer interaction (HCI), or machines networked together as in distributed artificial intelligence (DAI). Frequently, the principles are not rigidly thought out and explicitly expressed; they are implicit in systems and theories whose emphasis is elsewhere.

> " Previous DAI work has assumed that agents are mutually cooperative through their designer's fiat" (Rosenschein, 1985). Rosenschein offers as an example Lesser and Corkill's system (Lesser & Corkhill, 1983) in which agents are always assumed to be working on the same problem. "..it makes little sense to ask *why* they are helping one another; they help each other because they have been designed that way.." (Rosenschein, 1985).

The accepted notion of cooperation which is so implicitly encoded, involves agents firstly having a common goal. In some cases this may involve each agent in a multi-agent system individually holding a goal to achieve the same end result. In other cases, the goal may be initially held by only one of those agents but on recognition of this, other agent(s) then take on that goal as their own. Davis and Smith's (1983) "contract net" framework for example, comprises a range of problem-solving nodes. Each node, on receipt of a task divides it into subtasks, and offers these to the other nodes . Nodes bid for the subtasks and the original node then assigns these after examining the bids. This example demonstrates

the view of cooperation as sharing; the delegation or sharing of tasks between agents in complete agreement as to their goal(s). The negotiation as to how the goals are satisfied is merely in terms of node allocation. There is an assumption of total concordance concerning the existence of goals, and their division into subgoals.

Taking on another's goal as one's own implies an element of "helpfulness". Even if each agent individually and independently has the goal p, this "helpfulness" component must be present, if there is to be actual cooperation between them to achieve p. In other words, the reason for having the goal must be at least in part because another agent has it as a goal. Power refers to the situation of more than one agent independently having the same goal, and yet unaware of each other's goals, as "accidental coordination" (Power, 1984). He offers as an example two art enthusiasts who independently arrive at an art gallery with the goal of destroying a particularly shocking picture. One is intercepted by a guard, but the other succeeds in tearing the picture. The one agent achieved the goal state of both of them, and yet unaware of the other's goal, he was not actually cooperating to achieve it. If however, they had collectively possessed the intention to damage the picture, they would have had a mutual intention (Power, 1984).

Allwood also characterises cooperation as including a "mutual consideration" as well as a common purpose (Allwood, 1976). In fact, he goes further by defining mutual consideration as not merely an awareness of the existence of another agent's goal. That is just the cognitive consideration component. In addition there is an ethical consideration which states that agents should not do anything which would prevent one another from acting in accordance with the rules of rational agenthood as described in section 2.2.7. They should therefore not prevent each other from acting intentionally and purposefully according to their will, or from acting normally according to their motives, or from being rational. Allwood's definition of ideal cooperation for normal rational agents, of which communication is an example, can be summarised as : a number of individuals voluntarily striving to achieve the same purpose whilst ethically and cognitively considering each other in trying to achieve those purposes, and trusting each other to do this unless they give explicit notice to the contrary (Allwood, 1976).

Previous computational research in cooperative interactions has embodied the notion of helpfulness in cooperation as the recognition of another's goal. Agents are also designed to always be cooperative - always helpful and therefore ready to take on other agents' goals. This assumption is generally built in

to the system architecture. One agent merely needs to communicate her goal to another agent, for cooperation in the form of joint planning, to ensue. Examples include the systems described earlier from Davis & Smith (1983) and Lesser & Corkill (1983). Some more examples:

One of the earliest pieces of research to consider cooperation as an issue in conversation was Power's model of conversation involving joint plan formation. John and Mary were separate parts of a program who used conversation as a means of agreeing plans, exchanging information and so on eg. John announces a goal to Mary. He asks Mary to cooperate. She agrees. He then asks her to help him make a plan. She agrees....and so on. John and Mary's representations of the planning tree are always identical (Power, 1979).

More recent work by Shadbolt & Musson similarly involves joint planning to perform a task, in this case housebuilding. The agents need to cooperate because neither can build the house alone. If a completed task is a precondition to another task, and the agent cannot complete the first task herself, she requests the other agent to perform it. Being cooperative and helpful, on request, the task is performed (Shadbolt & Musson, 1986, 1987).

Cohen & Perrault's earlier work on speech acts, includes a "cause-to-want" "act". Its precondition is that the receiving agent believe that the agent has a goal. This is achieved by that agent issuing a request. The effect of a "cause-to-want" which therefore follows from a request, is that the receiving agent takes on that goal (Cohen & Perrault, 1979).

Appelt's system, KAMP, similarly assumes "if one agent is helpfully disposed towards another and knows that the other agent intends to bring something about, he then adopts that goal as his own." He adds: "

> It is seldom true that a person will want everything he knows that another person wants. However, if the domain of discourse is restricted to a cooperative endeavour [eg. the task in a task-oriented dialogue], this assumption will suffice to produce reasonable behaviour." (Appelt, 1985).

Finally, Rosenschein (1981) defines cooperativity as a predicate: COOP (x,g,t) whereby :

"x is "cooperative" towards y at time t in the sense of adopting as his own goals whatever he perceives y's goals to be" (Rosenschein, 1981).

Is there any condition, any circumstance, in which helpful agents in existing research, do not necessarily adopt other agents' goals by virtue of merely recognising their existence? Yes. Cohen and Levesque define agents as helpful in adopting others' goals, but only as long as they do not have any existing contradictory goal of their own (Cohen & Levesque, 1987b). Perrault also stipulates this condition (Perrault, 1987). In fact, this condition is most probably true in all the other work mentioned as well. However, it is not explicitly stated because the implicit assumption is that agents simply **do not have conflicting goals of their own.**

There are only a few examples of previous work in artificial intelligence, which acknowledge the existence of conflicting goals between cooperative agents. Firstly, the work of Georgeff which concerns the synchronisation of robots with existing plans. However, he uses a "supervisor process" which enables potential conflicts to be successfully avoided; they are not dealt with directly (Georgeff, 1983). The work of Rosenschein, actually focuses on the role of conflict in everyday encounters. He considers that cooperative agents must be able to deal with conflict; it should not be assumed to simply not exist. His ideas and criticisms of existing artificial intelligence notions of cooperation, will be discussed in the following section (Rosenschein,1985, Rosenschein & Genesereth, 1985). Ginsberg (1987) tackles similar problems to Rosenschein and Genesereth, such as the problem of robot or agent cooperation in situations where the pursuit of a local goal actively discourages cooperation, in distributed systems. An example is the Prisoner's Dilemma Game. A decision procedure paradigm is employed which incorporates a variety of assumptions, one of which is common rationality. The role of communication is then to describe situations rather than intentions, and establish agreed payoff functions or utilities[1].

[1]The Prisoner's Dilemma Game and game theoretic notions such as utilities, are described in section 3.2.5.1.

3.2.2.2. Criticisms of existing notions of cooperation

The existing notion of cooperative multi-agent interaction, as outlined in the previous section can be summarised as follows: Firstly, there is a common goal between the interacting agents. Secondly, the agents are "helpful" in that at least a component of their possessing the goal is an awareness that another agent has it as a goal. Awareness of another's goal is sufficient to adopt that goal as one's own, assuming no contradictory one is already in existence. Finally, if the attainment of a goal requires the attainment of sub-goals, these are adopted and shared out between the participating agents.

There is nothing wrong with this as far as it goes. What it lacks is any mention of the possibility of conflicts. Is it the case then, that in order to cooperate, agents must be in full agreement? If so, is cooperation impossible in situations where there happens to be a conflict of interest? If this is also true, then how can cooperative systems designed on such a basis interact in real-world contexts, where the unpredictability of the environment requires flexibility of action? Can conflicts always be just avoided or ignored, or is it the case that **one important component of cooperative interaction is the joint resolution of conflicts?**

Rosenschein's thesis "Rational interaction: cooperation among intelligent agents", is aimed towards the design of machines for DAI which can plan cooperative action, yet acknowledging the existence of multi-agent conflict. The premises from which he starts are firstly, that "the vast majority of real-world interactions lie between the two extremes of total conflict and absence of conflict", and if machines are being designed to operate in the real-world, then they cannot "operate under restrictive, crippling assumptions" (Rosenschein, 1985). Secondly, given the unpredictable nature of real-world domains, performing well requires "sophisticated and flexibile interaction capabilities" (Rosenschein, 1985).

Agents being assumed to have identical goals, and carrying out any task requested of them, has been termed by Rosenschein "the benevolent agent assumption" (Rosenschein & Genesereth, 1985). Whilst discussing how previous work in DAI operates within this assumption, he offers scenarios for which "intelligent" machines are currently being envisaged, and for which greater flexibility would be essential. One of these is that of an automated personal secretary interacting with another automated or human secretary, in order to schedule a meeting. In order to be practical, the private interests of the human boss must be protected.

"The full complexities of this scenario (commitments, threats, quid pro quo offers, hiding of information) require fully flexible interacting agents, beyond the capabilities currently addressed in AI systems." (Rosenschein, 1985).

Another example scenario concerns robots constructing a space station, motivated fundamentally by the same goal.

"...in the course of construction, however, there may be minor conflicts caused by occurrences that cannot be fully predicted (eg., fuel running low, drifting of objects in space). The building agents, each with a different task, could then negotiate with one another and resolve conflict" (Rosenschein & Genesereth, 1985).

In other words, Rosenschein is suggesting that previous systems are embodiments not of cooperative assumptions regarding the nature of agents, but assumptions of benevolence. Cooperation is in fact more than helping others in the absence of conflict. Cooperation can also result from the resolution of conflicts, and this should be an expectation of systems designed for real-world application.

"By allowing conflict of interest interactions, we can address the question of why rational agents would choose to cooperate with one another, and how they might coordinate their actions (even without communication) so as to bring about mutually preferred outcomes." (Rosenschein,1985).

Durfee, Lesser & Corkill (1987) consider that Rosenschein and Genesereth misleadingly refer to agents who share goals, as benevolent. They agree that "Benevolence is neither assumed nor needed for the agents to cooperate" (Durfee, Lesser & Corkill, 1987). Their argument however, is that cooperation should be viewed by agents as in their self-interest. With improved communication resulting in agents in a distributed network having a good level of common knowledge of network activity, each can decide about actions which will enable the network as a whole to behave more coherently. The motivation for such action is not imposed benevolence, but mutually advantageous action being in each node's

self-interest. My view of this approach is that it is in fact the benevolence assumption in another guise. Agents still always cooperate, but this time according to an understanding of the benefits from doing so. There is therefore no acknowledgement of the role that conflict can play in the maintenance and evolution of cooperation in changing and unpredictable environments. Conflicts are programmed out of this automated and artificial environment; they are necessarily detrimental to the system.

In the theory of multi-agent interaction proposed in this thesis, conflict is not a "dirty" word. Rosenschein has observed the "ubiquity of conflict" in multi-agent interactions, and argued the necessity for corresponding flexibility of action in an unpredictable and changing world. This research now takes these ideas a little further. Incorporated into the proposed theory of multi-agent interaction upon which the generation and interpretation of dialogue actions is based, is the notion that **conflict not only exists between agents, but plays a positive role in the maintenance and evolution of cooperation.** It has an important role in the maintenance and stability of cooperative multi-agent systems. This claim is based on a theory of social conflict and social change. A description of this is given in Chapter 5, together with justifications for its relevance to cooperative multi-agent systems comprising distributed artificial intelligence as well as human-computer interaction.

The conclusion here is that: **this theory of multi-agent interaction considers that agents operating cooperatively in the real world, must have a realistic understanding of cooperation. This involves understanding the nature and role of multi-agent conflict.**

> "..no group can be entirely harmonious, for it would then be devoid of process and structure. Groups require disharmony as well as harmony, dissociation as well as association; and conflicts within them are by no means altogether disruptive factors." (Coser, 1956).

To conclude this section: the definition of cooperation incorporated into this theory includes a common goal which may be generated as a result of recognising it as another's goal, but this is conditional upon the agent's own preferences. Helpfulness is retained as an important element of cooperation by commitment to the common goal being relative to the other agent having it as a goal. It is the inclusion of preference which so importantly removes benevolence from the definition, and

replaces it with agent autonomy. This definition is discussed further in section 3.2.4 and chapter 5.

3.2.3 Dialogue and posture - the nature of negotiation

It was said in the previous section, that agents with explicit representations of the three propositional postures, conflict, cooperation and indifference, described in terms of patterns of mental states, could potentially manipulate any one of these three states to be any of the others. What is needed in addition, is a rationality whereby agents believe themselves and each other to be able to effect such changes. What are the means of effecting changes to agents' mental states? Following from what was said in section 3.2, where the actions of agents were described as the cause of changes to the state of the multi-agent system, dialogue actions as a special case of actions in general (Austin, 1962) are the means of effecting changes to the mental states of the interacting agents. Therefore, for example, an agent x with a goal to have lunch with y, for which she obviously needs y's cooperation, may recognise that y has an existing contradictory goal. If she wants cooperation, she must alter either her own or y's mental states so the pattern of mental states between x and y in relation to the proposition concerning x and y lunching together, corresponds to that for cooperation. A component of this is that there must be a common goal; either x drops her goal or y eventually has a goal to have lunch with x. y can potentially take on x's goal if she recognises its existence, and if she has a preference of her own which would result in x's goal.

If x understands the effects of dialogue action on mental states and therefore on posture, one way in which the posture can be altered to one of cooperation is firstly to inform y of x's goal. Secondly, she can attempt to manipulate y's perception of the situation to one where y sees x's goal as preferable to her existing contradictory one. If successful, y would revise her goal. The conditions for this, as described in section 2.2.2.1 and chapter 6, would be satisfied. All x has to do is to generate a belief in y that she is faced with two alternatives for which she has a particular preference, and from which therefore goals can be generated and dropped. For x this involves reasoning with her own beliefs about what y's preferences might be. An example may be x offering to take y to her favourite restaurant.

Power writes that:

"there is something unnaturally Machiavellian about the idea that the essence of conversation consists in the manipulation of the partner's mental state" (Power, 1986).

My answer to this is that the sinister aspect is not in the fact of manipulating mental states. Any appearance as such is dependent upon the assumed nature of agenthood and agent rationality. For example, in artificial intelligence research to date, agent x having a goal p, and manipulating y's mental states as a means of achieving p, is an example of purposive action. However, purposive action is surely more than maximising personal gain, simply in terms of p. Agents rarely have individual goals. As explained in chapter 2, agents have multiple goals, and some of these are more central to the agent than others. The goal to never knowingly do harm to anyone may be a goal much more central and resistant to change than the goal to persuade someone to one's own point of view. An agent with strong beliefs and goals concerning the "common good" would plan to effect changes to others' mental states according to these beliefs and goals. Purposive action may conform to a collective rationality and therefore be that which maximises joint gain. Individual rationality is concerned with the maximisation of purely personal gain[1].

Power's alternative is a theory of collective planning to be applied to conversation (Power, 1986). All goals are joint goals, and problems such as a dispute over the means to achieve them, would be taken as a planning problem, solvable by solutions such as backtracking to try and find an alternative action which is not problematic, for example. As with the research described in sections 3.2.2.1 and 3.2.2.2, by avoiding conflict this ignores its positive role in cooperation.

The expression of a conflict can result in a cooperative solution by potentially dealing with the problem and removing it. Avoidance of conflict however, is side-stepping the issue. Dialogue has the potential to actually alter the circumstances which comprise the conflict. Dialogue alters mental states; conflict and cooperation are patterns of mental states. **This research focusses on dialogue as a means of potentially removing conflict, in order to achieve cooperation.**

[1]The terms "collective" and "individual" rationality have been adopted from game theory. They are discussed in section 3.2.5.1.

"Negotiation" is the term used to refer to this process of removal of the conflict, or reaching agreement. It is defined as:

> "conferring with another with the purpose of securing agreement on some matter of common interest" (Morley, 1977).

The interpretation of negotiation used in this thesis includes agreement about issues; not merely about the allocation or delegation of issues with assumed lack of disagreement on the issues themselves, such as in Davis & Smith (1983).

Reaching agreements is a process which occurs in many cooperative contexts. There is increasing experimental evidence which indicates that cooperative interactions such as doctor and patient (Cacciari 1985) or expert and advice-seeker (Kidd 1985) involve elements of negotiation. This supports the point made in the previous section that conflict is an important component of the maintained cooperation and stability in social systems. For example, when consulting an expert, people frequently have ideas about possible solutions to their problem. Taking advice is not merely a matter of being given the right answer. Previous misconceptions have to be satisfactorily dispelled, adequate explanation given as to the reasons why one solution is more effective than another, and so on.

In summary, the property of interacting agents which has been discussed here is: **agents believe themselves and others to be able to use dialogue actions to effect changes to the posture of the system, by effecting changes to the beliefs and goals of other agents. An example is agents believing themselves and others to be able to negotiate.**

Rosenschein's work is more concerned with action in general than with dialogue as a means of achieving cooperation, but he does consider the role of promises and deals (Rosenschein, 1985). In fact, from similar initial premises about conflict resolution and flexible cooperative action for real world applications, Rosenschein offers quite a different solution to that offered here:

Single actions are chosen in the light of an existing conflict situation. Reasoning about the most appropriate action is according to game- theoretic principles of maximising the payoff (some background to game theory can be found in section 3.2.5.1). Interaction strategies are analysed according to four alternative rationality principles and assumptions as to what types of move the opponent will make. In

other words, conflict is not avoided or removed; it is managed in terms of finding a "best action" in the circumstances. The role of communication is explored in terms of agents with conflicting goals making binding promises or deals, as to future action. The role of communication for information passing is also considered, but only between agents with common goals. The agents have to converge upon a particular plan of action and may have discrepancies between facts in their databases. The support and justification for these facts is compared in the light of various heuristics, and assuming a non-individualistic basis for the justification.

Rosenschein's work offers a choice of "best action" in the light of existing circumstances; this thesis suggests negotiation - the use of dialogue action to remove conflicts and change the circumstances. It is acknowledged that both approaches are important and valid. In some situations, the conflict may be such that no amount of negotiation results in agreement; action must simply be taken. Perhaps communication becomes impossible for some reason. A possible area for future research could concern agents' assessments of whether negotiation is or has become inappropriate to the eventual attainment of cooperation. Rosenschein's suggestion is that cooperative solutions to conflict encounters which do not involve communication, exploit the computational aspects of a computation/ communication trade-off. They are efficient means of dealing with conflicts (Rosenschein, 1985).

> "..certain definitions of rationality will imply certain types of cooperation, even in conflict-filled, communication-free, non-benevolent interactions" (Rosenschein, 1985).

3.2.4 The autonomy of interacting agents - control of information

If we take on board that cooperation is more than agents doing whatever is asked of them because they are assumed to never disagree, then interacting agents no longer have the property of benevolence. Agents may have goals as a result of what is asked of them, but these should also be the result of their own preferences in the matter. Such agents are autonomous. If agents assume themselves and others to be autonomous in this way, then each agent has only partial control over the effects of their dialogue actions. For example, A asks B to make a cup of tea. This is a purposive action and is understood as

such by B. In other words, both A and B believe A to have a goal for some tea. However, although the goal for tea was the cause of the request made by A, A only has partial control over its effect and the goal being satisfied. B as an autonomous agent may take on A's goal as his own and make the tea, or he may not.

The conditions for belief and goal adoption were described in chapter 2, as evidence (in the case of belief) and /or preference. B recognises A's goal and is faced with the alternatives of making the tea or not making the tea. His preference concerning these exact circumstances is the cause of the goal which is generated. If B has an existing contradictory goal, then B will adopt A's goal only if on recognition of A's goal, B prefers it to his current goal and its consequences. This in turn rests on the other goals he has and how fundamental they are. Perhaps helping people whenever possible may be a goal which can only be satisfied once A has made the request, by B deciding to make the tea. This is embodied in B's preferences.

The conclusion is that when reasoning about purposive action, agents must take into account this autonomous nature of themselves and others. Just because a particular dialogue action is one which could potentially generate a specific belief or goal in another agent and thus satisfy a goal, no assumptions can be necessarily made as to the effect of the action being as desired. I can want you to believe that I'm funny by telling you a joke, for example. You may not be convinced. Receiving agents therefore have control over their own mental states - they control the information they acquire. Acting agents know this.

In summary: **Agents believe themselves and others to be able to <u>potentially</u> manipulate desired changes to other's mental states, and therefore also to the posture of the system, using dialogue actions. This is conditional upon either evidence in the world or the preferences of the receiving agent.**

If dialogue actions are potential manipulators of others' mental states, then they can also be manipulators of other agents access to one's own mental states. In other words, just as receiving agents have control over the mental states they acquire, acting agents have control over the information they reveal. In this theory, agents are not assumed to be sincere or otherwise. They perform speech

acts which can be veracious or mendacious, revealing or concealing expressions of mental states. Veracity, mendacity, concealing and revealing are defined (in chapter 6) as properties of acts, which reflect a relation between the act and a mental state of the actor.

The motivation for providing definitions of this sort is an extension of the original motivations concerning the acknowledgement of the role of conflict in cooperation. The notion of cooperation here has already been described; just as cooperative agents are not assumed to be benevolently helpful, neither are they assumed to be benevolently sincere. Dialogue as strategic interaction is introduced in the next section, and strategies can be various. It is important to remember however, that people have beliefs about the potential effects of using these different expressions. For automated agents operating in the real world, these should also be an important component in the ascertainment of preferences and goals, and therefore in reasoning as to their use:

> "Everyone depends on deception to get out of a scrape, to save face, to avoid hurting the feelings of others. Some use it more consciously to manipulate and gain ascendancy. Yet all are intimately aware of the threat lies can pose, the suffering they can bring. .....Those who learn that they have been lied to in an important matter - say, the identity of their parents, the affection of their spouse, or the integrity of their government - are resentful, disappointed and suspicious. They feel wronged; they are wary of new overtures" (Bok, 1978).

> "In lying to others we end up lying to ourselves. We deny the importance of an event, or a person, and thus deprive ourselves of a part of our lives. Or we use one piece of the past or present to screen out another. Thus we lose faith even with our own lives" (Rich, 1975).

Rosenschein refers to "potentially catastrophic side-effects" which are not easily discovered, from the passing of incorrect information (Rosenschein, 1985).

However, Goffman claims that everyday interaction involves elements of control of information. "Expression games" are the individual's capacity for acquiring, revealing and concealing information. He draws on popular literature on intelligence and espionage, because:

"just as we are like them [wanted criminals, spies and secret police] in significant ways, so they are like us....occasions are always arising when we must ask for advice and then determine how to read the advice by trying to analyze the sincerity of the server's manner. When we come into contact with the person who employs us, a similar task arises; he has reason to almost cover his actual assessment with an equable, supporting air , and we have reason to try to read this for what he "really" thinks." (Goffman, 1970).

Bok also refers to "*levels* of deception that we must all live with.....all around have clustered the many kinds of deception intended to mislead." (Bok, 1978). She refers to changes of subject, disguises, evasive or exaggerating gestures:

"all blending into the background of silence and inaction only sometimes intended to mislead. We lead our lives amidst all these forms of duplicity." (Bok, 1978).

It therefore seems appropriate to include in this theory, that **agents believe themselves and others to be able to choose to perform dialogue actions which are mendacious and/or concealing expressions of a mental state, <u>but</u> agents are assumed to use revealing and veracious expressions unless the receiving agent has any beliefs which might suggest the contrary.**

"There is no general law against lying. Yet there is a marked tendency for people to tell the truth, as they see it, at least when the lie offers no conspicuous gain." We "expect a built-in tendency towards veracity on the part of the speakers and towards credultiy on the part of the listeners" (Quine, 1970).

"Lying requires a *reason*, while truth-telling does not" (Bok, 1978).

In contrast to the above, Cohen and Levesque offer sincerity as a property of agents. Communicative acts are successfully performed when the speaker is assumed to be sincere.

"We describe agents as sincere and helpful. Essentially, these concepts capture (quite simplistic) constraints on influencing someone else's beliefs and goals, and on adopting the beliefs and goals of someone else as one's own. More refined versions are certainly desirable. Ultimately, we expect such properties of cooperative agents, as embedded in a theory of rational interaction, to provide formal descriptions of the kinds of conversational behaviour Grice (1975) describes with his "conversational maxims"" (Cohen & Levesque, 1987b).

In this thesis' framework, communicative actions are assumed to be generated on the basis of the speaker believing it to be a good strategy. This means that if someone asks you to "Go jump in a lake", then you interpret it literally as a request if you believe they have a goal for you to jump in a lake and believe you will adopt this as an autonomous agent, on the basis of your own preferences. Your subsequent action will also be in their favour. In situations where there is no lake around, or the conversation until this point had nothing to do with lakes, the speaker is still believed to have generated the utterance with a communicative goal that was believed by them to be a good strategy. There must therefore be an alternative goal which you are intended to recognise, adopt and respond favourably to. This is much like Grice's maxims of quality, quantity, relevance and manner, which if apparently flouted are in fact being conformed to at some level. For this framework however, there is only one such principle. Agents don't need to be themselves formally represented as having separate properties describing them as being sincere, relevant, informative, or unambiguous. Good strategy is one general principle which the utterances of negotiation dialogues between rational, autonomous agents are assumed to conform to. It is used together with the principle expressed above concerning agents use of revealing, veracious expression unless there is reason to believe the contrary. Just as Cohen and Levesque concentrate on the property of sincerity, this principle reflects the focus of interest of this research. It could be extended however, to encompass the other maxims such that agents are assumed to use relevant, unambiguous, informative expressions unless there is reason to believe the contrary. Apparently ambiguous or irrelevant utterances are then interpreted during negotiations, as with the "Go jump in the lake" example above, according to the principle of good strategy.

The precise nature of good strategies, and more detailed discussion of all these issues including a

comparative analysis with Cohen and Levesque's approach, can be found in chapter 6 . A description of the nature of strategic interaction follows in the next section.

3.2.5 Strategic interaction

Negotiation is considered an example of a type of dialogue and of strategic interaction. For this, agents require a strategic rationality. The role of the rest of the sections comprising section 3.2.5 is to describe these terms, deriving them from their background in game theory and psychology.

3.2.5.1 The relevance of game theory

The game-theory definition of a game of strategy is as follows: interdependent decision making in social situations between two or more autonomous agents, each of which has only partial control over the outcomes (Colman, 1982). This conforms to the views expressed in this theory of multi-agent interaction, whereby the performer of dialogue actions believes themself and others in the multi-agent system to be autonomous. This means that the outcomes of a dialogue action, which are effects on the mental states of the participating agents, cannot be determined by the speaker alone. **Dialogue is a game of strategy.**

The study of games of strategy or game theory, is based upon von Neumann and Morgenstern's utility theory (von Neumann & Morgenstern, 1944). Each possible outcome for an interaction in a particular situation is assigned a numerical utility. These utilities reflect relative preference which:

> "may be based on any factor whatsoever that influence the player's degrees of satisfaction or dissatisfaction with the possible outcomes, including spiteful or altruistic attitudes towards the other player(s), religious beliefs, phobias, masochistic tendencies, and so forth" (Colman, 1982).

The action selected is the one which maximises the expected numerical utility; rational action conforms

to preferred outcome.

The games analysed are either of pure conflict, total cooperation or mixtures of conflict and cooperation - mixed-motive interactions. Does such an approach have anything to offer this theory of cooperative multi-agent interaction, as applied to negotiation dialogue?

To analyse negotiation in game theory terms would require generating a mathematical model of all possible outcomes to alternative dialogue actions. Each would have utilities expressing a subjective and consistent relative preference for the speaker. This would relate to the numbers of goals it satisfies, and a subjective probability of the successful attainment of the desired mental state. This is the SEU model, or subjectively expected utility maximisation model, which became popular amongst game theorists in 1954 (Edwards, 1967). The rational move is the one which then provides the maximum payoff. But, rational for whom? The term "rational" needs additional clarification; it is not clearly defined in game theory.

> "Loosely, it seems to include any assumption one makes about the players maximising something" (Luce & Raiffa, 1957)

or:

> "Roughly speaking, rationality is concerned with the selection of preferred behaviour alternatives in terms of some system of values, whereby the consequences of behaviour can be evaluated" (Simon, 1959).

> "Perhaps the only way to avoid, or clarify, these complexities is to use the term "rational" in conjunction with appropriate adverbs.....A decision is "organizationally" rational if it is oriented to the organization's goals; it is "personally" rational if it is oriented to the individual's goals" (Simon, 1959).

The maximum payoff for the individual in the dialogue situation described above would be the rational choice, if the agent is operating according to an individual rationality. On the other hand, if there is the possibility of agents operating themselves and/or believing others to be operating according

to a collective rationality, then the optimum choice of action could be the one where the joint or collective gain is the greatest. This is applicable to negotiation as an example of a bargaining game, which is a special case of the n-person cooperative game. In these, the result or outcome has to be mutually agreed, and players can form subsets, termed coalitions. In a two person game, the four possible coalitions are either no players, player one alone, player two alone, or both players together. Coalitions offer their members a better payoff than they would get by operating individually and independently. The classic Prisoner's Dilemma Game is an example of a two-person game where a coalition between the players could benefit both. There is one pareto-optimal outcome which is preferred by all, because it balances the payoffs of the players in a collective gain. The possibility of greatest individual gain in this particular game however, is by non-cooperative action.

The consideration of another player possibly operating according to a collective rationality is especially relevant to personal choice of action in dialogue. This is because it is not a single move game. An individual's action is often strongly affected by what they expect the other player to do next (Colman, 1982). In other words, the expected consequences of each of the possible dialogue actions, in terms of subsequent action on the part of the other player(s), should be a component of the comparative process. Howard developed a theory of metagames in order to allow for just this aspect of interaction (Howard, 1971, 1974); players choose from a selection of metastrategies which are dependent on predictions of the other(s)' action. It is:

> "the game that would exist if one of the players chose his strategy after the others and in knowledge of their choices" (Howard, 1971).

For example, two players each have two possible strategies in the basic game, A and B. The first level metagame is constructed by player 1 with four possible metastrategies, each dependent on which of the two strategies player 2 might choose next. These are: AB (do A if player 2 will choose B), AA, BA, and BB. The second level metagame offers player 2 ,16 possible metastrategies, - four metastrategies each dependent on player 1's choice out of four. (An example analysis using this method can be found in section 7.5.3).

The importance of Howard's work is that in considering the other player's subsequent options, the

assumed rationality is not necessarily individualistic. In this way, the metagame approach offers a cooperative solution to the classic game, the Prisoner's Dilemma Game, for example. Choosing to conceal important information which potentially results in both players being released by the police, is a metarational move. In classic game theory, the only rational choice in this game is individualistic - disclose the information with the possibility of personal reward as well as release, at the cost of the other player's punishment and imprisonment. At worst both can then only be imprisoned, but still without punishment (Colman, 1982, Howard, 1971).

Metagame analysis allows choices of action to be made on the basis of predictions of the opponent's behaviour which cover all options. These include the collective, even the altruistic, as well as the individualistic. Richelson has applied this approach to the study of strategic deception between the United States and Soviet Union. He demonstrates that deception in terms of misrepresentation of their attitudes concerning commitments to conventional and nuclear attack, is advantageous (Richelson, 1979). However, being a game analysis and specifically about threat of nuclear attack, Richelson's study does not take into account any further consequences of the deception, such as those unrelated to the one interaction under study, but with implications for future dealings. Therefore, the kinds of constraints to deception referred to in section 3.2.4, are not taken into consideration. For this reason, his work only has any relevance to particular kinds of dialogues where the interacting parties know they will never again come into contact with each other, or anyone else that may know them.

There are many more question-marks regarding the relevance of game-theoretic analysis to social situations involving decision making in the real world. It has been claimed to serve as "a skeletal analogy of many social situations and contexts", and yet this is refuted by others who say that games and real conflict situations do not correspond (Schlenker & Bonoma, 1978). Games are static situations involving straightforward interactions between players, with clearly defined, consistent and frequently minimal goals. Practical contexts on the other hand, involve complex agents each with multiple goals of different strengths, which may change during the course of the interaction. The limits of computational resources are also a point to be considered; searching for optimal strategies through several thousand possible outcomes is possibly a waste of effort.

One attempt at a less purely mathematical and more psychologically plausible approach to game theory was introduced in 1977 by Bennett. He developed a theory of hypergames, these being a set of

perceptual games expressing each individual player's perspective (Bennett, 1980). His objective was to offer an alternative game theoretic analysis which did not require the players to be perfectly and correctly informed of each other's strategies and preferences.

A metagame analysis which was carried out on an example dialogue early in this research programme, is described in section 7.5.3. This and other game approaches were eventually abandoned however. Partly this was due to the questions of applicability to real-world conflict situations, briefly referred to above. More details on these developmental stages in the research programme can be found in section 7.3. However, many useful ideas concerning the nature of choice and strategic interaction developed out of investigations into this area.

3.2.5.2 Strategic rationality

Howard's and Bennett's work, described in the previous section, are particular forms of expression of some of the ideas also claimed by both the game theoretician, Schelling, 1960, and the psychologist, Goffman, 1970. These have been incorporated into this theory of multi-agent interaction, and concern the incorporation of a strategic rationality for dialogues of the negotiation type. This can be summarised as:

in negotiations, autonomous, rational agents choose dialogue actions on the basis of firstly, the mental state they want to induce in the hearer, secondly, whether they believe this mental state will be achieved, and thirdly, subjective expectations of subsequent dialogue action on the part of the hearer also being in the speaker's interests.

The first two elements of this are embodied formally within the definition of INTERESTS, described in section 4.4.4.2. Interests were introduced in section 2.2.5 as a means of representing goals which the agent believes will be achieved. This is stronger than merely an achievable or feasible goal, and is intended as a means of incorporating a notion similar to the subjective probability of success element in

SEU's. Probabilities, however, can be graded precisely and relatively, according to numerical values assigned by applying some algorithm. Such a mathematical approach is inapplicable to the current means of expression adopted by this framework. This practical issue is discussed further in section 4.4.4.2.

The third element of the definition of strategic rationality says that: rational and strategic dialogue action is consistent with the speaker's expectations that subsequent dialogue action on the part of the hearer, will be in the speaker's interests. This is therefore the component which concerns predictions as to the other agents actions. These expectations are based upon beliefs about the other agent's actions also conforming to a strategic, and either individual or collective rationality. This involves reference to beliefs about previous encounters in other social situations of a similar type, and perhaps with the same agents. According to Goffman, such predictions are made by:

> " directly orienting oneself to the other parties and giving weight to their situation as they would seem to see it, including their giving weight to one's own ... An exchange of moves made on the basis of this kind of orientation to self and others can be called strategic interaction." (Goffman, 1970).

Goffman offers this in addition to Schelling's definition concerning strategic interaction as action which constrains the future actions of others to one's own advantage:

> "A strategic move is one that influences the other person's choice, in a manner favorable to oneself, by affecting the other person's expectations on how one's self will behave. One constrains the partner's choice by constraining one's own behaviour." (Schelling, 1960).

This thesis claims verbal strategic interaction to be dialogue action performed by multi-agents in accordance with the definition of strategic rationality given earlier. Firstly, it is consistent with the goal state which is the mental state immediately desired to be induced in another, and this is one which the agent believes will be achieved. Secondly, it is consistent with the believed future action resulting as a consequence of that mental state being achieved, also being desired by the speaker. Therefore the goal

comprising the first phase is potentially a means of inducing a particular action on the part of the hearer, which indirectly achieves another goal of the speaker. Its role is to constrain the opponent's subsequent choices of action in one's own favour. For example, two children, A and B are arguing over a doll. B suggests to A that she could have his sweets if she gives him the doll.

> "The object is to set up for one's self and communicate persuasively to the other player a mode of behaviour ... that leaves the other a simple maximisation problem whose solution for him is the optimum for oneself.." (Schelling, 1960).

In the example above, B believes A to prefer sweets to the doll. From this he predicts that subsequent to his offer, A will lose interest in the doll and give it up. This is because he believes that she believes that on recognising her action, B will then give up his sweets. Therefore, in addition to Schelling's notion of strategic interaction, the application of this definition of strategic rationality also incorporates Goffman's ideas concerning mutual assessments of each other. The basis for action is a prediction about the other agent, based on beliefs about her preferences, and also her beliefs about oneself. These ideas are formally expressed as a predicate Gd-STRAT, and incorporated into the definition of a communicative act, for strategic interaction in chapter 6.

It is important to point out, that reasoning strategically about interaction may not be useful in all dialogues; perhaps only those where:

> "each party must make a move and where every possible move carries fateful implications for all of the parties" (Goffman, 1970).

This point has in fact already been made in chapter 2 in relation to the processing effort required to determine prior to action, whether a goal state is believed to be eventually achieved and therefore in one's interests, as opposed to just being achievable. Similarly, considering future actions as a component of reasoning about the action from which these would result, is computationally costly, and only worth doing if the consequences of not doing it are severe. Having said this, it is acknowledged that there should be some means of assessing when and whether such thorough reasoning is necessary.

Negotiation was defined in section 3.2.3 simply as "conferring with another with the purpose of securing agreement on some matter of common interest" (Morley, 1977). This embodies both trivial and non-trivial disputes, either of which could have inconsequential or significant potential implications for the involved parties. The fact that negotiation and the employment of deceptions play a part in everyday conversation as well as in more obvious contexts such as political disputes, has already been discussed in sections 3.2.3 and 3.2.4. Full-blown strategic reasoning is unlikely to be appropriate for all exchanges, comprising all such cases. However, the basis upon which its appropriateness can be determined is a refinement which will have to wait for future research.

3.3 Conclusion

Multi-agents are interdependent components of a system of relationships. Of particular interest to this research are the postural relationships of conflict, cooperation and indifference, and these are characterised as patterns of mental states. Conflict is viewed as serving a positive function in the maintenance and stability of cooperative systems.

Multi-agents understand the nature of posture. They understand the role and conditions for dialogue to effect postural changes, whilst taking into account individual agents' autonomy over their mental states, and therefore the flow of information in the system.

An example of multi-agents potentially effecting postural change is negotiation. This is a type of dialogue and strategic interaction, the objective of which is to attain a posture of cooperation with respect to some proposition, where previously the relation had been conflict or indifference. Strategic interaction conforms to a strategic rationality. For interaction using dialogue, this means that dialogue actions are consistent with the speaker's goals and expectations of success, as well as favourable expectations concerning any subsequent action on the part of the hearer. Such dialogue actions are strategic tools; they are the means by which agents can achieve either their own and/or joint goals. Thus they also conform to either an individual or collective rationality.

The next chapter describes the formal means of expression for these ideas. The rest of the thesis then comprises the details of the major points of this theory of cooperative multi-agents, and how they fit

together into a formal framework with the theories of agenthood and speech acts as described in previous chapters.

CHAPTER 4 : A formal approach to modelling theories of agents

4.1 Introduction

The theories of agents, multi-agents, and cooperative interaction described in the previous chapters, need to be expressed in a form which demonstrates and tests them as a basis upon which speech actions can be pragmatically generated and interpreted. This after all, is the major theoretical premise; utterances are generated and interpreted in context, that context including importantly, the mental states of the speaker and hearer. This mental context includes all relevant attitudes concerning the nature of multi-agent interaction, which alongside the immediate goals of the interaction, determine action. The concern of this thesis is to evaluate the particular principles of cooperative interaction proposed as a basis for dialogue, and such dialogue as a means of maintaining and evolving cooperation in multi-agent systems.

The theoretical principles can be expressed as explicit properties; properties of individual agents, properties of acting agents, properties of interacting multi-agents, properties of cooperative, interacting multi-agents. With these, agents are accessing the principles of rational and cooperative interaction with which to reason about dialogue action. They have a model of what it is to be a rational agent cooperatively interacting in a multi-agent system.

The first requirement from which such a model of multi-agent interaction can be developed, is a model of what it is to be an individual rational agent in the world. This should conform to the theoretical conditions described in chapter 2. A formalism has in fact been devised by Cohen and Levesque (1987a) specifically as a means of writing specifications of rational agents. Similarly to the aims of this research, this has then been demonstrated by them to be a basis for communication between cooperative agents (Cohen & Levesque, 1987b). The formalism and some aspects of their model of agents have been adopted here.

The role of this chapter is to describe the formalism and corresponding formal representation of agents. Section 4.4 is devoted to this following which, formal representation of the proposed principles of cooperative multi-agent interaction are then detailed in chapters 5 and 6. A brief survey of the theoretical background, context and comparison with alternative formal approaches and derived models of

agents, is given in section 4.3. Section 4.2 concerns methodological issues. Firstly, an affirmation with justification of the primary theoretical concerns - these being the nature of the problem of cooperative dialogue, and not the mechanistic details of exactly how this should be physically realised in future systems. Secondly, in the light of this, the reasons for the particular choice of representation and reasoning system which has been used here as a means of demonstrating, testing and evaluating the theory.

4.2 Why a formal approach?

> "Artificial Intelligence is the study of complex information-processing problems that often have their roots in some aspect of biological information processing. The goal of the subject is to identify interesting and solvable information-processing problems, and solve them" (Marr, 1981).

Marr's view is that there are three levels at which such information-processing problems can be considered. The first two concern the nature of the problem; an abstract formulation of what is being computed and why, which he refers to as the computational theory. Finally, there is how this is achieved. Such a separation of specification and implementation concerns is in fact quite a general approach; the specification and description of a software engineering problem prior to its implementation, for example. It is the computational theory which Marr considers to be most important. Because the same algorithm can frequently be implemented in many different ways, it is therefore better explained in terms of the nature of the problem it solves than the mechanisms it uses to do so. He cites Chomsky's competence theory for English syntax as an example of a computational theory which is:

> "little concerned with the gory details of algorithms that must be run to express the competence (i.e. to implement the computation)" (Marr, 1981).

The aim of this research is to develop a theoretical framework for computational models of cooperative dialogue, acknowledging the role of conflict in multi-agent cooperation. The

information-processing problem is the generation and interpretation of speech actions by autonomous and rational automated agents, in order to resolve conflicts and achieve cooperation. This requires reasoning with intensional concepts such as knowledge, belief, goals and intention. The algorithm to be explained is the reasoning processes whereby such mental states are transformed into speech actions (or vice versa for dialogue interpretation). This thesis therefore comprises a computational theory in Marr's terms, or a specification in software engineering terms. In order to describe as well as then test, evaluate and develop it, an appropriate means of expression has been chosen. This is a type of modal logic; mental states and actions are represented and reasoned with using modal operators and a possible-worlds semantics. It should be stressed however, that the role of this is solely as a means of expression or description of the theory or specification. Appropriate physical realisations in future computer implementations are a separate issue, and one only to be considered once the theory is established. It is also not the case that the use of a logic implies an emphasis on logical concerns.

The reasons for a logic being the chosen representation medium, concern its advantages as a means of expression. Logics are languages in which there are rigorously laid down *syntactic* and *semantic* requirements, and all expressions have to conform to these. Meanings and implications of expressions are consequently incontrovertible. In addition, there are no complicated control structures to confuse the central issues. A theory expressed as a set of logical axioms is evident; it is open to examination. This assists the process of determining whether any parts of the theory are inconsistent, or do not behave as had been anticipated when they were expressed in English.

The major points therefore are:

1) Logics are languages with precise semantics. A semantics determines what every expression in a language means, and logics are defined in terms of exact specifications for this. A consequence of this is that there can be no ambiguities of interpretation. In addition, the soundness of the inference procedure can therefore be checked (Reichgelt, forthcoming book). Every sentence derivable from the set of axioms of the logic should be a valid consequence of those axioms. The axioms used to describe the properties of beliefs for example, are therefore assumed to be consistent with the intended meaning of beliefs as portrayed by the model (Konolige,1986).

2) By expressing the properties of agents, and multi-agent systems as logical axioms and theorems in a language with a clear semantics, the focal points of this research are explicit. The theory is transparent; properties, interrelationships and inferences are open to examination. This contrasts with the use of computer code, which requires implementational and control aspects within which the issues to be tested can often become irretrievably enmeshed. It is frequently the case, that computer systems concerned with joint activities, such as problem-solving, are in fact designed such that properties of the interacting agents are implicit properties of the entire system, and it is impossible to investigate the role or effects of any individual aspect.

3) The particular logic chosen was designed by Cohen and Levesque (1987a, 1987b) specifically to specify the properties of cooperative rational interaction in a clear, explicit and unambiguous form. This work builds on their results, as well as questions some of their assumptions. The same formalism has been used for a clear comparison of results, and to be able to focus on the development of the theory rather than the creation of an alternative means of expression or implementation.

The negative aspects of this choice result from some fundamental problems with the particular model for belief representation upon which it is based. The two primary mental states beliefs and goals, which along with actions act as the 'input' and 'output' to the algorithm, are represented as modal operators with a possible-worlds semantics according to the Hintikka model (Hintikka, 1962). It is well known that this model represents agents as logically omniscient, "ideal knowers"; they not only believe all valid formulas, but believe all the logical consequences of their beliefs. By adapting this for goals, agents also have goals for all the consequences of any existing goals. Despite these difficulties however, the possible-worlds model is widely considered one of the most promising so far. Its shortcomings acknowledged, it has been adopted by cognitive scientists whose primary interests are in developing theories of agenthood and communication with whatever best means of representation is currently available. This is the spirit in which it has been adopted also here. Some of the research currently taking place to modify, develop or radically alter the possible-worlds approach towards a more realistic representation of belief and other mental states is referred to in section 4.3.1.2.

There are general arguments amongst researchers in artificial intelligence as to the validity of

employing logics of any sort for representation and reasoning systems. For example, there are arguments concerning psychological plausibility. Johnson-Laird (1986) quotes many current theories which embrace the philosophical belief that laws of thought are nothing but laws of logic. He claims that humans are indeed rational thinkers, but it is not logic which underlies this. This view is also associated with Marvin Minsky. He considers the correct approach to AI problems is to attempt to imitate the way the human mind works, and this is not by mathematical logic (Kolata, 1982). John McCarthy on the other hand, considers that:

"This is A(rtificial) I(ntelligence) and so we don't care if it's psychologically real" (Israel, 1983).

There is a place in AI for both of these viewpoints.

This thesis describes research which is unconcerned with developing a psychologically plausible model; the aim is the development of a framework whereby automated agents can produce the desired behaviour, regardless of whether they do it in the same way as humans. There is also no insistence on the use of logic as the implementation language in future implementations.

Another argument with respect to the use of logics in knowledge representation and reasoning concerns the emphasis therefore on deductive inference. McDermott's recent paper "A Critique of Pure Reason" considers that deductive inference alone is insufficient; there are two important types of reasoning, abduction and default reasoning, which are not deductive (Reichgelt, 1987). D. Israel (1983) also considers as too strong, the claim that all that is required are deductively valid rules of inference. "Logical proof is a tool used in reasoning." It should be "kept in its proper place" (Israel, 1983).

In Reichgelt (forthcoming book) the arguments in favour of logic are stated as being primarily the benefits of having a precise semantics as described earlier. There are also a variety of different logics available, such as temporal and epistemic logics. Epistemic logic and its role in the modelling of agents is described in more detail, in the next section.

4.3 Formal approaches to the modelling of agents

Having discussed the rationale behind the decision for a formal means of expression, this section describes the background and context of the particular choice of formalism. The existence of alternative models of agenthood arises initially out of differences in the representation of knowledge and belief.

4.3.1 Knowledge and belief representation

A very brief overview will be given in this section of the variety of formal systems which have been applied to the problem of representation of belief and knowledge. The objective is not to provide an in-depth comparative review, but merely to give enough information to indicate the possible alternatives to the adopted approach. Justification is given for the choice made, which is described in more detail in section 4.3.1.1.

Konolige (1986) distinguishes the different formal approaches to belief representation along two parameters:

1) the model of belief. The alternatives are either a symbol-processing or sentential model, versus beliefs as propositional attitudes.

> "...in the former, an agent's beliefs are characterized by the computations an agent performs on syntactic objects (symbol strings or sentences) in some internal language; in the latter, belief is taken to be a relation between an agent and abstract propositions about the world" (Konolige, 1986).

2) the language of formalisation. The syntactic approach considers belief as a predicate in a first-order metalanguage which express facts about sentences. A set of sentences comprises an internal or object language. The denotations of terms in the metalanguage are therefore expressions in the object language. The alternative is the use of belief as a modal expression, generally with a possible-worlds semantics.

Sentential, symbol-processing models are generally expressed with syntactic logics. Examples are found in Konolige (1982), Montague (1963), Perlis (1985). Konolige's deduction model however, is a sentential model formalised in a modal language (Konolige,1986). The propositional attitude model on the other hand, usually coincides with the use of a modal language and possible-worlds semantics. Examples of this are the systems of Hintikka (1962), McCarthy (1978), Sato (1976), Moore (1980), Levesque (1982), Halpern (Fagin & Halpern, 1985) and Vardi (Halpern, 1986); applications include studies on language understanding and communication by Appelt (1982, 1985) and Cohen and Levesque (1987a, 1987b). The advantage of syntactic languages is expressive power. There is full quantification over the sentences of the object language, whereas generally in modal languages there is only quantification over individuals. However, they are "notationally burdensome" (Konolige, 1986) because there must be a set of terms referring to expressions in the object language as well as terms for individuals and predicates.

Use of a modal logic and possible-worlds semantics has been by far the more popular approach. It is formally elegant in being:

> "representationally more compact, and doesn't suffer from the proliferation of confusing terms referring to object language expressions that the syntactic approach is prone to" (Konolige, 1986).

The major drawback of this approach however, is its assumption that agents are ideal knowers and ideal reasoners. They are "logically omniscient" and have infinite computational resources. Adaptations to the possible-worlds approach to weaken these assumptions are currently a popular research issue.

> "A number of attempts have been made to modify the possible-worlds framework to provide a more realistic semantic model of human reasoning ... While none of these attempts appears as yet to provide the definitive solution, they do suggest that there is sufficient flexibility in the possible-worlds approach to make it worth pursuing" (Halpern & Moses, 1985).

4.3.1.1 Modal logics and possible-worlds semantics - the "classical" model

The mental states of beliefs and goals are modelled in this research as propositional attitudes and expressed in a modal language using possible-worlds semantics. The details of the formal language devised by Cohen and Levesque (1987a, 1987b) and described in detail in section 4.4, are based upon the semantic approach to the representation of belief, now known as the "classical" model and which is briefly described in this section.

The use of possible-worlds semantics to reason about epistemic notions such as knowledge and belief, was an idea originally conceived by Hintikka (1962). Knowledge or belief are axiomatised as the necessity operator ($\Box$) in S5 modal logic. The axioms for strong S5 are as follows (weak S5 is the same but without axiom M3):

M1: P, where P is a tautology.

M2: $\Box(P \supset Q) \supset (\Box P \supset \Box Q)$

M3: $\Box P \supset P$

M4: $\Box P \supset \Box \Box P$

M5: $\sim \Box P \supset \Box \sim \Box P$

The inference rules for standard modal logics are modus ponens and necessitation. If P is a theorem then so is $\Box$ P. Then if $P \supset Q$ is a theorem, then so is $\Box P \supset \Box Q$. When related to knowledge or belief, these properties of the system make agents logically omniscient; if an agent believes P then she must also believe Q.

Hintikka adopted the possible-worlds semantics for modal logic which originally Kripke had constructed (Kripke, 1963). The meaning of the modal operators knowledge and belief, are defined in terms of accessibility relations among possible worlds. Reasoning therefore also concerns the relations between the different possible worlds in which a proposition holds, not just the truth of the proposition. The intuitive idea behind this is that for every true state of affairs, there are many other possible states of affairs. These can be described as possible worlds. An agent knows P, iff P is true in every world she thinks possible. An agent does not know P iff there is at least one possible world where P does not

hold. Halpern (1986) gives as an example of this an agent who believes two states of the world to be possible. In both it is sunny in San Francisco, whilst in one it is sunny in London and in the other it is raining in London. The agent knows it is sunny in San Francisco but does not know whether it is sunny or raining in London.

The propositions which are objects of an agent's beliefs comprise a set of possible worlds, and all of these are compatible with the agent's beliefs. Each of these worlds is accessible from the others via a belief relation. The propositions which are objects of an agent's knowledge, comprise a set of possible worlds, and all of these are compatible with what the agent's knows. These worlds are also accessible to each other but via the knowledge relation.

The properties of the operators knowledge and belief, extend from the properties of the accessibility relations. For example, if the real world has to be one of the possible worlds and therefore the knowledge relation is reflexive, then an agent cannot know anything false. This is expressed as axiom A3. or $Know_{\chi}P \supset P$ which is M3. as shown above, but with Know substituted for the necessity operator. The belief relation on the other hand, does not have this reflexive property and therefore this axiom is absent from the characterisation of belief. Beliefs are therefore endowed with the property of not necessarily relating to "truth" as existence in the real world. Both the knowledge and belief relations are also transitive and symmetric. Agents correspondingly have the property of knowing what they know and believing what they believe (see M4 above), and knowing what they don't know and believe (see M5 above) (Halpern & Moses, 1985).

4.3.1.2 Semantic approaches and logical omniscience

As mentioned earlier, the above "classical" possible-worlds approach to the representation of knowledge and belief results in models of agents who know and believe all valid formulas and know and believe all the logical consequences of their knowledge and beliefs. This is generally acknowledged as an unrealistic representation of the epistemic aspects of agenthood. It was also said earlier that the model adopted here makes use of Cohen and Levesque's formal approach (Cohen & Levesque, 1987a, 1987b), which being based upon the "classical" model, therefore suffers from these problems.

Although the emphasis of this research has been clearly stated as not concerning representational

issues, it is important to mention just a few examples of work by logicians aimed at developing modifications of this model, or alternatives. It is obviously crucial that such developments occur in order that there eventually be realistic and applicable implementations of computational theories concerning information-processing problems which require reasoning with beliefs.

Halpern (1986) discusses some of the most notable attempts at solving especially the problem of logical omniscience. One of these is to augment standard possible worlds with "impossible" worlds" where the customary rules of logic do not hold. This approach has not been widely adopted however, because nonintuitive semantic rules are used to assign truth values to the logical connectives, and "it is not clear to what extent this approach has been successful in truly capturing our intuitions about knowledge and belief" (Fagin & Halpern, 1985).

Levesque (1984) adopts a different, more "computationally attractive" (Levesque, 1984) and "intuitively plausible" (Halpern, 1986, Fagin & Halpern, 1985) semantic approach to the representation of belief, than the "classical" possible-worlds model by distinguishing between implicit and explicit belief.

> "... a sentence is *explicitly* believed when it is actively held to be true by an agent and *implicitly* believed when it follows from what is believed..." (Levesque, 1984).

Implicit beliefs therefore have the characteristics of beliefs in the classical model; they consist of all the logical consequences of what is explicitly believed. A new semantics for the explicit belief operator B, is provided by replacing possible-worlds by partial possible-worlds or situations, and a three-valued truth function. Explicit belief is identified with a set of situations rather than possible worlds, and a situation may support the truth of some sentences, the falsity of others, and may not deal with some sentences at all. Levesque's situations are essentially those of situation semantics (Barwise & Perry, 1983).

Fagin & Halpern (1985) reinterpret Levesque's explicit belief in terms of limited awareness on the part of the agent of some propositions. "How can someone say that he knows or doesn't know about p if p is a concept he is completely unaware of?" (Fagin & Halpern, 1985). They also offer an alternative logic which in addition to the modal operators B and L for explicit and implicit belief, includes an awareness operator A. Explicit knowledge consists of implicit knowledge plus awareness; an agent

explicitly believes p if p is true in all worlds the agent believes possible and p is in the agent's awareness set.

A quite different approach is Rosenschein's situated-automata approach to modelling knowledge. This models the logical relationships between the state of a process, referred to as a machine, and that of its environment. There are constraints between a process and its environment which mean that not every state of the process-environment pair is possible. A process knows a proposition p, in a situation where its internal state is s, if in all possible situations in which the process is in state s, p is satisfied. This definition satisfies the axioms of S5 modal logic, including deductive closure and positive and negative introspection, but does not require the encoding of sentences of a formal language as data structures. It is extended to hierarchically constructed machines; the agent or robot as a machine which comprises individual components - elements of a multi-agent system. This model of knowledge reasons about the flow of information between the machine's components. It avoids inferential complexity yet provides a concrete computational model of knowledge which allows real-time performance (Rosenschein & Kaelbling, 1986). Halpern (1986) points out that an essentially identical notion of knowledge as distributed amongst processors and described in terms of the states of each of these, was developed independently by himself and Moses (Halpern & Moses, 1986), and others working in distributed systems. Their emphasis was not computational, however.

It is interesting to note that the representation of agents as societies of multi-agents, is an approach which has simultaneously found favour in very different research areas; in robotics and distributed artificial intelligence as has just been described for example, and Fagin uses a similar notion in his logic for "local reasoning". Fagin & Halpern (1985) suggest a logic in which agents are "societies of minds". Each has its own cluster of beliefs which may contradict each other. This allows "local reasoning" whereby agents can hold inconsistent beliefs, focussed upon in different "frames of mind" (Fagin & Halpern, 1985). The idea is also associated with Minsky (1981).

Yet another alternative approach treats propositions as sets of worlds, and knowledge and belief as sets of propositions for each agent (Vardi,1986).

4.3.2 Moore's model - reasoning about knowledge and action

The motivation for the volume of research into the representation of knowledge and belief, is that it is an initial stage towards modelling the relationship between these attitudes and action. Moore's work (Moore, 1980) was very influential in devising a formalism which allowed explicit reasoning about aspects of this relationship, described by Halpern as follows:

> "Knowledge is necessary to perform actions, and new knowledge is gained as a result of performing actions" (Halpern, 1986).

Reasoning about the ability to perform actions based upon a general understanding of this relationship between knowledge and action, enables plans to be carried out in the light of incomplete information. For example, a plan can be made to open a safe when the combination is unknown, if the agent understands that there is possibly some action by which this knowledge can be determined, such as reading the combination from a piece of paper.

Moore constructed a logic in which the effects of action are described in terms of a modal logic parallel to the modal logic for knowledge. The two logics are unified by identifying the situations in the semantics of the logics of action with possible worlds in the semantics of knowledge. The effects of actions on knowledge are described in terms of relations between possible worlds. A modal operator, RES, is introduced for this, which is parallel to KNOW for knowledge. The semantics for RES are in terms of an accessibility relation, R, exactly as K is for knowledge.

Knowledge is axiomatised as an S4 modal logic, not S5 as was described for the Hintikka schema in section 4.3.1.1. In Moore's system therefore, M5 is excluded from the accessibility relation K, and consequently agents do not know what they do not know.

To describe the effects of an action on an agent's knowledge, the set of possible worlds compatible with the agent's knowledge both before and after performing the action are described. This is because, what is possible according to the agents knowledge after performing an action is the result of performing the action in some world that was possible according to her knowledge, before performing the action.

Appelt (1985) points out that Moore's use of possible worlds to represent situations, which are states of the world resulting from the performance of an action, is an unorthodox interpretation. Traditionally, possible-worlds are an entire course of events. They include a temporal history of events and the truth of propositions are defined at each point in time. Cohen and Levesque's treatment of possible worlds described in section 4.4, is in accordance with this (Cohen & Levesque, 1987a, 1987b). Moore's alternative approach uses possible worlds in which the truth of all propositions are defined at a single instant of time, and therefore reasoning about temporal relations involves reasoning about sequences of possible worlds (Appelt, 1985).

Appelt (1982, 1985) adopted Moore's work as the basis of the reasoning component of his planner KAMP. This plans the generation of utterances. Moore's system uses a reified approach whereby knowledge is axiomatised as a modal operator, and the possible-worlds semantics for the modal logic is then axiomatised in a first-order logic. Reasoning by automatic deduction is therefore possible, using a conventional theorem prover for first-order logic. Appelt's system made use of the practicality of this; basic facts about objects, relations, actions, and mental states are stated in an intensional object language but reasoned with only after translation into a first-order metalanguage. The benefits of maintaining the modal language in spite of its need to be translated, are that it is concise and comprehensible, and can be used to derive concepts which cannot be derived in the possible-worlds, first-order language.

Both Appelt and Moore only consider the relation between knowledge and action. Beliefs are a weaker notion than knowledge in that they may be false; plans may fail as a consequence of false belief. Although recognising the relevance of this to realistic application, especially in natural language generation, Appelt chose to adopt Moore's line and restrict the analysis to knowledge whilst concentrating on other issues.

4.3.3 Cohen and Levesque's model - reasoning about intention and commitment

In contrast to the model described above, Cohen and Levesque's work focuses more on the relationships between beliefs and goals, and action, than knowledge and action (Cohen & Levesque, 1987a, 1987b). Belief is axiomatised as a weak S5 modal logic, as was described for the "classical"

model. This is adapted for goals. A goal is a desired state of the world; what the world would be like if the agent's goal was true. This includes the world the agent is currently in. In other words, agents cannot want what they currently believe to be false; goals are a subset of beliefs. Their treatment of goals is slightly unusual in this respect; mostly a goal is considered to be a future-directed attitude. Cohen and Levesque differentiate between achievement or A-GOALs, and maintenance or M-GOALs, but in fact only use A-GOALs in their analyses. Their treatment of beliefs and goals suffers from the problems referred to earlier of necessitation and logical omniscience. They do not adopt Moore's treatment of possible worlds as situations. Their possible worlds are courses of events, extending infinitely into the past and infinitely into the future. Action expressions are included in the language, and their denotations events, are primitive entities. They can therefore reason with action, time, beliefs and goals. With this facility, Cohen and Levesque have developed a theory of agents with which to examine complex interrelationships and notions such as commitment and intention. By adopting their approach and model, this research extends this venture to examining other important issues of cooperative multi-agent dialogue.

Cohen and Levesque have developed a theory of rational interaction, based upon aspects of agenthood which are derived from interrelationships between the primitive concepts of beliefs, goals and actions. Examples of such aspects are committed goals, and intentions. They make explicit the conditions under which agents drop their goals, and therefore provide a notion of what it is to be committed to a goal. Intention is then defined as a commitment to perform an action, with properties such as that the intention is believed achievable, and that the agent need not intend all the expected side-effects of the intention.

> "Thus, even using a possible-worlds approach, one can get a fine-grained modal operator that satisfies many desirable properties of a model of intention" (Cohen & Levesque, 1987a).

Their formalism comprises an atomic layer consisting of the analysis of beliefs, goals and action upon which there is a molecular layer of new concepts defined out of these primitives. Using all of this, they erect a theory of rational interaction and communication from which properties of communicative

acts can be derived (Cohen & Levesque, 1987b). The theory of cooperative multi-agent interaction proposed by this thesis is developed similarly. Properties of an example communicative act are also derived and used for comparison with Cohen and Levesque's theory of rational interaction in chapter 6.

4.4 A formal model of rational agenthood

Cohen and Levesque's formalism and some important properties of beliefs, goals and actions have been adopted wholesale by this research. In addition, their notions of committed goals and intention have been taken on. What differs here extends from the introduction of a notion of preference as a fundamental and pragmatic determiner of goals - especially in connection with belief revision in dialogue, interests as another type of goal, strategic rationality as a feasible approach to cooperative interaction, and representations of alternative postures in the system. The theory of rational interaction thereby created and the overall framework for cooperative dialogue action is in accordance with the views expressed in chapter 3, and will unfold in detail throughout chapters 5 and 6.

The agent model is expressed in the logic devised by Cohen and Levesque which has a model theory based on possible-worlds semantics. There are four primary modal operators: BELief, GOAL, HAPPENS and DONE. With these, the relationships between agents beliefs, goals and actions are characterised. The temporal and action-related aspects of the model, are provided by the properties of HAPPENS, DONE and $\Diamond$. Agent attitudes are cognitive and conative and represented with BELief and GOAL. The model for these operators is the "classical" possible-worlds model for beliefs, assuming the Hintikka axiom schemata. This is adapted to deal with goals, and events are included as primitive entities to enable quantification over action. Of a set of possible worlds, each consists of a sequence or course of events that characterises what has happened and what will happen. Each possible world in the set of all possible worlds is modelled as a linear course of events extending infinitely into the past and infinitely into the future. Consistency of some worlds with agents beliefs and goals is specified by an accessibility relation on worlds, agents and an index into the course of events that defines the world.

The syntax and semantics of the logic are described in detail in sections 4.4.1 and 4.4.2. The properties of Cohen and Levesque's model which have been taken on are detailed in section 4.4.3.

Section 4.4.4 contains some of the properties original to this research which are crucial to the definitions given in later chapters, of the focal aspects of cooperative multi-agent interaction[1].

4.4.1 Syntax

The logic uses predicates and existential quantifiers.

Variables are :

a,b.... variables ranging over acts and

x,y... agent variables.

Predicates are:

(<Predicate-symbol> $<variable_1>$ ..... $<variable_n>$)

Wffs are:

<Predicate> | ~<wff> | <wff> ∨ <wff> | one of the following:

$<variable>_1 = <variable>_2$

∃ <variable> <wff> where <wff> contains a free occurrence of variable <variable>.

(HAPPENS <action-expression>) - action-expression happens next.

(DONE <action-expression>) - action-expression has just happened.

(AGT x a) - x is the only agent of action a.

(BEL x <wff>) - <wff> follows from x's beliefs.

(GOAL x <wff>) - <wff> follows from agent x's goals.

[1]All theorems, propositions, definitions and assumptions with "C&L" in their name are taken directly from Cohen and Levesque (Cohen and Levesque, 1987a).

$(a \leq b)$ - a is a subsequence of b.

Action expressions are:

a | one of the following:

<action-expression> ; <action-expression> - sequential action

<wff> ? - test action

<action-expression>* - iterative action

4.4.2 Semantics

Model Theory:

A model M is a structure $< \otimes, P, E, Agt, T, B, G, \Phi >$, where $\otimes$ is a set of things, P is a set of people, E is a set of primitive event types, $Agt \in [E \rightarrow P]$ specifies the agent of an event, $T \subseteq [Z \rightarrow E]$ is a set of possible courses of events (or worlds) specified as a function from the integers to elements of E, $B \subseteq T \times P \times Z \times T$ is the belief accessibility relation, and $G \subseteq T \times P \times Z \times T$ is the goal accessibility relation. Formulas are evaluated according to some possible world and an "index" into that world, n.

D is the domain of quantification. $D = \otimes \cup P \cup E^*$ where E^* denotes sequences of events from E.

$\Phi \in [\ Pred^n \times T \rightarrow 2^{Dn}\ ]$, specifying the interpretation of predicates.

$AGT \subseteq T \times P$, where $x \in AGT[e_1, \ldots\ldots e_n]$ iff there is an i such that $x = Agt(e_i)$. Agt specifies the partial agents of a sequence of events.

Satisfaction

Assume M is a model, σ is a sequence of events, n an integer, v a set of bindings of variables to objects in D, and if $v \in [Vars \rightarrow D]$, then v^{x}_{d} is a function which yields d for x and is otherwise the same as v.

1. M, σ,v ,n $\models$ P (x_1,.....,x_n) iff < v (x_1), ... v (x_n), > $\in$ Φ [P, σ,n]. The interpretation of predicates depends on the world σ, and the index into it, n.

2. M, σ,v ,n $\models$ ~ α iff M, σ,v ,n $\not\models$ α

3. M, σ,v ,n $\models$ ($\alpha \vee \beta$) iff M, σ,v ,n $\models$ α or M, σ,v ,n $\models$ β

4. M, σ,v ,n $\models$ ($x_1 = x_2$) iff v (x_1) = v (x_2)

5. M, σ,v ,n $\models$ $\exists$ x α iff M, σ,v^{x}_{d} ,n $\models$ α for some d in D.

6. M, σ,v ,n $\models$ (AGT x_1 e_2) iff AGT [v (e_2)] = { v (x_1)}. AGT specifies the only agent of event e_2.

7. M, σ,v ,n $\models$ <Time proposition> iff v (<Time proposition>) = n.

8. M, σ,v ,n $\models$ (BEL x α) iff for all σ^* such that $\sigma B[v (x)]\sigma^*$, M, σ^*, v ,n $\models$ α. That is α follows from the agent's beliefs iff α is true in all possible worlds accessible via B, at index n.

9. M, σ,v ,n $\models$ (GOAL x α) iff for all σ^* such that $\sigma G[v (x)]\sigma^*$, M, σ^*, v ,n $\models$ α. That is α follows from the agent's goals iff α is true in all possible worlds accessible via G, at index n.

10. M, σ,v ,n $\models$ (HAPPENS a) iff $\exists$m, $m \geq n$, such that M, σ,v ,n [[a]] m. a is a sequence of events that happens next after n. [[]] $\subseteq$ [Tx Z x D x $Action\ expressions$ x Z]. It relates an action expression to two indices on a course of events.

11. M, s,v ,n |= (DONE a) iff ∃m, m ≥n , such that M, s,v ,m [[a]] n .

[[]] ⊆[*Tx Z x D x Action expressions x Z*] is characterised by:

event variables : M, σ,v ,n [[x]] n +m iff v (x) = $e_1 e_2 ... e_m$ and

$$\sigma(n + i) = e_i, 1 \geq i \geq m.$$

null actions : M, σ,v ,n [[NIL]] n

Alternative actions: M, σ,v ,n [[a|b]] σ_1 iff M, σ,v ,n [[a]] σ_1 or

M, σ,v ,n [[b]] σ_1

Sequential actions: M, σ,v ,n [[a;b]]m iff ∃ k, n ≤ k ≤ m, such that

M, σ,v ,n [[a]] k and M, σ,v ,k [[b]] m

Test actions: M, σ,v ,n [[α ?]] n iff M, σ,v ,n |= α

Iterative actions: M, σ,v ,n [[a*]] m iff ∃n_1,.....,n_k where n_1 = n and n_k = m and

$$\forall i,\ 1 \leq i \leq m \quad M, \sigma, v, n_i [[a]] n_{i+1}$$

Some important abbreviations:

Conditional action :

(IF α THEN a ELSE b) =*def* α?; a | ~α?;b

- as in dynamic logic, an if-then-else action is a disjunctive action of doing action a at a time at which α is true or doing b at a time at which α is false.

Eventually:

◊α = *def* ∃x (HAPPENS x; α?)

- ◊α is true if α is true some time in the future - there is something that happens after which α holds.

Always : $\Box\alpha$ = *def* $\forall x$(HAPPENS x) $\supset$ (HAPPENS x; α?)

$\Box\alpha$ means α is true throughout the course of events.

Constraints on the Model:

The constraint on G is that it is contained in B. This means that chosen worlds are a subset of the agents beliefs. $\forall \sigma, \sigma^*$, if $< \sigma, n >$ G [p] σ^*, then $< \sigma, n >$ B [p] σ^*. In other words, B $\supset$ G. This is a "realism" constraint which says that an agent only chooses worlds which are included in the set of worlds the agent believes to be possible (C&L, 1987).

From $G \subseteq T \times P \times Z \times T$, for a given agent at a given time-index in a given world, G will return a set of worlds which the agent would choose as satisfying its goals. This set is then a subset of the worlds which would be returned by B for that agent at the given time-index in the given world, as the worlds the agent considers it could be in , according to its beliefs.

B is Euclidean, transitive and serial for any agent x and time-index, n. B being Euclidean means that the worlds accessible from any world via (B x n) form an equivalence class, but not necessarily including the world the agent is in i.e. the real world. G is serial i.e. there is always a world which is accessible to the given world via the B- and G- relations.

4.4.3. Properties taken directly from Cohen and Levesque's model

4.4.3.1.Properties of acts and temporal modalities

C&L Def 1. (DONE x a) =*def* (DONE a) $\wedge$ (AGT x a)

x is the agent of the act a, which is in the past i.e. the act has been done.

C&L Def 2. (HAPPENS x a) =*def* (HAPPENS a) $\wedge$ (AGT x a)

An action a occurs, and x is the agent of that act.

(HAPPENS a) says an action occurs.

(HAPPENS ~p?; a; p?) says that event a brings about p.

C&L Def 3. (LATER p) = *def* $\sim p \wedge \Diamond p$

p is not true now, but will become true in the future.

For $\Diamond$ read "eventually". $\Diamond p$ is true iff somewhere in the future, p becomes true. $\Box p \equiv \sim\Diamond\sim p$. $\Diamond p$ and $\sim\Diamond p$ are jointly satisfiable.

C &L Proposition 5: $\models p \supset \Diamond p$

C&L Proposition 6: $\models \Diamond(p \vee q) \wedge \Box\sim q \supset \Diamond p$

If eventually either p or q is true and q is forever false , then eventually p.

C&L: Proposition 7: $\models \Box (p \supset q) \wedge \Diamond p \supset \Diamond q$

If p implies q at all times, and eventually p, then eventually q.

C&L Def 4:

(BEFORE p q) =*def* $\forall$c (HAPPENS c; q?) $\supset \exists$ a($a \leq c$) $\wedge$ (HAPPENS a; p?)

p comes before q if, whenever q is true in a course of events, p has been true.

4.4.3.2. Properties of attitudes

Beliefs:

The following axiom schemata follows from the Hintikka characterisation of knowledge (Halpern and Moses 1985). It corresponds to a "weak S5" modal logic. Axiom A3 is missing from the characterisation of belief; it states that only true facts can be known.

A1. All instances of propositional tautologies.

A2. $(BEL\ x\ p) \wedge (BEL\ x\ (p \supset q)) \supset (BEL\ x\ q)$ - an agents' beliefs are closed under implication.

A4. $(BEL\ x\ p) \supset (BEL\ x\ (BEL\ x\ p))$ - positive introspection i.e. an agent has beliefs about what she believes.

A5. $\sim(BEL\ x\ p) \supset (BEL\ x\ \sim(BEL\ x\ p))$ - negative introspection i.e. an agent has beliefs about what she does not believe.

A6. $(BEL\ x\ p) \supset \sim(BEL\ x\ \sim p)$ - consistency of beliefs.

R1. $\dfrac{p \wedge (p \supset q)}{q}$ - modus ponens.

R2. $\dfrac{p}{(BEL\ x\ p)}$ - necessitation.

N.B. R2 and A2. make agents "ideal knowers" i.e. not only do they believe all valid formulas, but they also believe all the logical consequences of their beliefs. This is the logical omniscience problem.

C&L Def 5: (KNOW x p) = *def* p ∧ (BEL x p)

An agent knows p if the agent believes p and p is true.

C&L Def 6: (COMPETENT x p) = *def* (BEL x p) ⊃ (KNOW x p)

Agents competent with respect to p have beliefs about p which are true. Agents are assumed to be competent with respect to their own beliefs, goals, their having done primitive events.

C&L Def 7: (ABEL n x y p) = *def* $\underbrace{(\text{BEL x (BEL y (BEL x ...}}_{n}\underbrace{\text{(BEL x p) ...)}}_{n}$

ABEL characterises the nth alternating belief between x and y that p, built up from the "outside in".

C&L Def 8: ((BMB x y p) = *def* ∀ n, (ABEL n x y p)

BMB is the infinite conjunction of :

(BEL x p) ∧ (BEL x (BEL y p)) ∧ (BEL x (BEL y (BEL x p)))).....

Goals:

Goals are consistent i.e.

C&L Proposition 16: |= ~ (GOAL x False)

They are also closed under consequence:

C&L Proposition 17: |= (GOAL x p) ∧ (GOAL x p ⊃ q) ⊃ (GOAL x q)

There is a necessitation property:

C&L Proposition 18: If $\models \alpha$ then $\models (\text{GOAL } x\ \alpha)$

Agents eventually drop their goals:

C&L Assumption 3: $\models \Diamond\sim(\text{GOAL } x\ (\text{LATER } p))$

C&L Proposition 3: $\models (\text{BEL } x\ p) \supset (\text{GOAL } x\ p)$

Agents do not want what they currently believe to be false. This means that if x has a goal for p to be true sometime in the future, then she does not believe it to be forever false, or impossible:

$(\text{GOAL } x\ \Diamond p) \supset \sim (\text{BEL } x\ \Box\ \sim p)$

C&L Def 9:

$$(\text{P-GOAL } x\ p) = \text{def}\ (\text{GOAL } x\ (\text{LATER } p)) \wedge (\text{BEL } x\ \sim p) \wedge$$
$$[\text{BEFORE } ((\text{BEL } x\ p) \vee (\text{BEL } x\ \Box\ \sim p))\quad \sim(\text{GOAL } x\ (\text{LATER } p))]$$

Agents have commitment to some goals. These are defined as persistent goals which are only given up if the agent believes it is achieved, or believes it is impossible to achieve. Cohen and Levesque refer to this as fanatical commitment.

C&L Def 10:

$$(\text{P-R-GOAL } x\ p\ q) = \text{def}\ (\text{GOAL } x\ (\text{LATER } p)) \wedge (\text{BEL } x\ \sim p) \wedge$$
$$(\text{BEFORE } [(\text{BEL } x\ p) \vee (\text{BEL } x\ \Box\ \sim p) \vee (\text{BEL } x\ \sim q)\ ]$$
$$\sim(\text{GOAL } x\ (\text{LATER } p)))$$

To remove the fanaticism from commitment, persistent goals can be defined such that they are relativised. These are goals which are only given up if the agent believes it achieved, believes it is impossible to achieve or the reason for the goal q, is false.

C&L:Theorem 2:

$$\forall y\ (\text{P-R-GOAL}\ y\ p\ q) \wedge \sim[\text{BEFORE}\ \sim(\text{COMPETENT}\ y\ p) \vee (\text{BEL}\ y\ \Box\sim p) \vee (\text{BEL}\ y\ \sim q)\ \sim(\text{GOAL}\ y\ (\text{LATER}\ p))] \supset \Diamond p$$

If someone has a persistent goal to bring about p relative to q, and before dropping her goal, p remains within her area of competence and the agent will not believe p will never occur or does not believe q to be false, then eventually p becomes true. Proved using :

C&L: Proposition 23: $\models (\text{P-GOAL}\ x\ q) \supset \Diamond[(\text{BEL}\ x\ q) \vee (\text{BEL}\ x\ \Box\sim q)]$

If the agent has a persistent goal that q, then she eventually either believes it is true or impossible to achieve, and C&L: Proposition 6.

Intention:

From the above, Cohen and Levesque have defined an intention to act:

C&L Def. 11:

$$(\text{INTEND}_1\ x\ a\ q) = \textit{def}\ (\text{P-R-GOAL}\ x\ [(\text{DONE}\ x\ (\text{HAPPENS}\ x\ a))?;\ a]\ q)$$

An intention to perform an action a, is defined as a persistent goal for a to happen next, but relative to some condition q.

4.4.4 Some additional properties of the model

In this section, the notions of agents having preferences and interests is defined. Preference is a crucial component of the property of autonomy over mental states and belief revision during dialogue, discussed in detail in chapter 6. Considerations of autonomy are important components of assessments of good strategy in dialogue, when interacting according to a strategic rationality. So are interests, and details of these strategic aspects of the framework are given in chapter 6. This section therefore

introduces some important properties of the agent model, in preparation for detailed analyses of their roles in the entire framework, in later chapters.

Preference and interests are molecular concepts, defined in terms of the primitive elements of belief and goal, in just the same way as Cohen and Levesque have defined intention and commitment.

4.4.4.1 Preference

Def 1: $(\text{PREFER } x\ p\ q) =_{def}$

$$(\text{BEL } x\ [\ (\text{BEL } x\ \Diamond(p \vee q)) \supset (\text{GOAL } x\ \Diamond p) \wedge \sim(\text{GOAL } x\ \Diamond q)]\ )$$

This says that an agent preferring p to q is defined as: the agent having a belief that if she could believe either that p or q will be true in the future, then she would have a goal that eventually p. She would not have a goal for eventually q. For example, agent x preferring to go out than stay in under certain conditions believes that if these conditions should prevail and she could believe either that she eventually will have gone out or have stayed in, then she would have a goal to eventually have gone out.

Assumption 1:

$$(\text{BEL } x\ \Diamond(p \vee q)) \wedge (\text{PREFER } x\ p\ q) \supset (\text{GOAL } x\ \Diamond p) \wedge \sim(\text{GOAL } x\ \Diamond q)$$

This says that, if the agent has a belief that if either p or q is eventually to be true and she prefers p to q, then she has a goal for p to eventually be true and does not have a goal for eventually q.

For example, if I prefer to believe "I'm pretty" to believing that "I'm ugly" this means that I hold a belief that if faced with these two options for the future, I'll generate a goal that I will believe "I'm pretty", and I won't therefore believe "I'm ugly". If in addition I am actually faced with the two options, then my preference leads me to generate the goal that I eventually believe "I'm pretty".

Proposition 1:

$$(\text{PREFER } x\ p\ q) \wedge (\text{BEL } x\ (p \supset r)) \supset (\text{PREFER } x\ r\ q)$$

preferring p to q means that x also prefers all the logical consequences of p eg r, to q.

As mentioned in chapter 2, agents are assumed to possess the machinery to be able to compute preferences, and a psychological theory for this was offered in section 2.2.3. It was hoped to be able to define preference as a function which evaluates specific preferences pragmatically according to this theory when required, but this proved very difficult. Given the focus of the research as the development of a theoretical framework for dialogue, I decided it best not to allow myself to get too side-tracked into this issue. The definition provided is adequate to the objectives of this research, which in accordance with previous related research such as Cohen and Levesque's (1987b), Moore's (1980) and Appelt's (1982, 1985), concerns the use of attitudes such as preferences, beliefs , goals and intention in dialogue, and not their determination. In this respect, this representation of preference is no better and no worse than that of beliefs and goals. Hopefully, there will be future research into the determination of attitudes.

The problems encountered included the necessity for ordering in relation to worlds which comprise an infinite set of propositions; comparisons between believing p and believing q for example, according to the numbers of consistent beliefs and goals which each sustains. Maximisation in relation to subsets of specifically relevant beliefs and goals would seem more appropriate, and Shoham uses this idea of subsets in his work on a general framework for nonmonotonic logics (Shoham, 1987). He introduces preference logics whereby standard logics are associated with a preference relation on models. Nonmonotonicity is the focussing on a subset of the interpretations or models that satisfy a formula, which are preferable in a certain respect. He therefore has introduced a partial ordering on interpretations and looks into the different possible preference criteria. However, the psychological basis for this in relation to ascribing a particular subset of beliefs and goals as those most relevant to a comparison with another such subset, in order to determine preferred beliefs or goals would be yet another research programme entirely. In addition, different beliefs and goals were described in section 2.2.2 as not being equivalent in their strength or centrality to the agent. Therefore, if believing p satisfies only one relevant goal and believing q satisfies three different relevant goals, the agent still may prefer to believe p to q, if that one goal is a very important and fundamental one. "Weightings" can be attached to each

belief and goal in terms of relative importance and the eventual algorithm take not only numbers of satisfied, relevant beliefs and goals into account, but their relative importance. However, the participating agents' mental states are changing throughout the conversation, and thus the relative value of each belief and goal in relation to each other one would constantly need to be reassessed.

> "Since changes in the certainty of one belief can exert seemingly arbitrary influence on the certainty of any other belief held by the agent, the problem of maintaining consistency of belief is very difficult" (Appelt, 1985).

In order not to allow these difficulties to undermine the testing of the theoretical intuitions regarding the role of preference in this strategic approach to cooperative dialogue, the examples in chapter 7 incorporate the assumption that agents understand the psychological basis for preferences as it is described in chapter 2. Preferences enter the framework as the end product of this assumed reasoning. Agents are therefore formally represented as having a set of preferences, just as they have existing sets of beliefs and goals. The examples in chapter 7 include demonstrations of how dialogue is used as a means of manipulating other's mental states, thus manipulating the context in which different of these preferences are relevant.

4.4.4.2 Interests

Def 2: (INTERESTS x p) = *def* (GOAL x (LATER p)) $\wedge$ (BEL x $\Diamond$p)

p following from x's interests is defined as x having a goal for p to become true in the future and believing that this will be the case.

For example, it is in x's interests to ask y to go out for a drink, if x has the goal that y go out with her, and believes that this will eventually happen. The basis for believing that it will eventually happen is according to the conditions for goal adoption in dialogue as mentioned in section 2.2.2.1 and detailed in chapter 6. x should believe y would prefer to go out with her than to not go out with her,

on recognition of her goal. This means that it doesn't matter if x believes y to have an existing goal to be going out with z. As long as her belief about y's preferences in this matter, lead her to believe she will adopt the goal, it is in x's interests to make the request.

In this example, it may seem unnecessary for x to consider if she believes y really will adopt her goal before making the request, as long as it is not believed impossible. The role of interests however is as a determiner of good strategy in strategic interaction. In certain contexts, it is important that rational action involves reasoning not only with one's goals, but with the believed likelihood of their successful attainment, and from this, expectations of any subsequent response. An example is the use of a lie where its discovery may have negative consequences. Interests are therefore intended as a means of incorporating a subjective determination of success into reasoning about strategically rational action. As explained in section 3.2.5.2, mathematical models for strategic action such as game theory or decision theory, use numerical probabilities for this. This model on the other hand, has no determiner of probabilities; the agent either believes her goal will be successfully attained, or otherwise. There can be no comparison between goals the agent may believe more likely to be satisfied than others. The assessments of likelihood and relative likelihood of successful goal attainment would be valuable developments, but ones which await further research in the formal expression of such notions. The problems are stated by McCarthy and Hayes:

> "1. It is not clear how to attach probabilities to statements containing quantifiers in such a way that corresponds to the amount of conviction that people have.
> 2. The information necessary to assign numerical probabilities is not ordinarily available. Therefore a formalism that required numerical probabilities would be epistemologically inadequate" (McCarthy & Hayes, 1969).

Halpern and McAllester (1984) propose a modal operator L for "likely" to allow qualitative reasoning about likelihood without the use of numbers. By adding a modal operator for knowledge there can be simultaneous reasoning about both knowledge and likelihood. They offer their logic LLK as a first step to being able to reason with these concepts, whilst acknowledging that more work needs to be done, to allow for statements such as "p is more likely than q", for example.

The role of interests in both the determination of desired mental states for induction in others, and good strategy is explained further in chapter 6.

4.5 Conclusions

This chapter has described alternative means of representing or modelling agents. Reasons are given for the particular approach considered the most suitable for this research. An agent model has been formally described appropriately to the theoretical discussions in chapter 2, regarding specifications for rational agenthood. In the following two chapters it is used as a formal basis with which to express the theory of cooperative multi-agent interaction introduced in chapter 3.

The agent model so described, is limited. It is limited firstly as a consequence of technical issues - those problems discussed in the body of this chapter, and associated with the particular formal means of representation. As a result, agents are modelled with idealised rationalities, demanding global consistency as opposed to reasoning with subsets of "local" beliefs, and having infinite sets of beliefs and goals each undistinguished according to the agent's awareness of them. Such agents would have problems communicating in complex environments, and certainly in anything approaching real time. The model is also limited however, as a consequence of theories of agents in cognitive science and artificial intelliegnce, being still underdeveloped. Some of the issues were discussed in chapter 2. These included a lack of acknowledgement of the role of affective attitudes in agent architectures, the restriction of conative attitudes to goals with little theory concerning the generation or determination of these except as sub-goals to existing ones, and the nature of rationality/ies.

As explained in section 4.2, the aim of this research is the development of a computational theory of dialogue. Current theories of agenthood are extended and developed into a theory of cooperative multi-agent interaction as a basis for the generation and interpretation of utterances. Representation is a secondary concern; its role is as a means of explicitly and unambigously expressing the theoretical intuitions, from which they can then be tested, and the theory correspondingly evaluated. Obviously however, the theory of dialogue which has been developed has had to accomodate to a certain extent, the limitations of its means of expression and component theories. This chapter has described early stages

in the representation of agents. New developments and insights in both the areas of agent theory and formal representation of agent architectures, are to be looked forward to in the future.

CHAPTER 5: Conflict, cooperation and indifference

5.1 Introduction

This chapter is devoted to discussions of the nature and role in multi-agent systems, of the propositional postures, conflict, cooperation and indifference. They are formally defined in the terms of this framework, with reference and comparison to ideas generated by other research. The use of postures as an essential element in strategic reasoning about dialogue action is demonstrated in chapter 7.

Posture was described in section 3.2.2 as a collective term denoting alternative characterisations of the social concepts, conflict, cooperation and indifference. Each is defined according to a pattern of mental states, and as a different relation between an agent, and another agent and a proposition. The definitions of conflict and cooperation both specify the nature and distinctions between the agent's own attitudes to a proposition and the attitudes believed to be held by the other agent with respect to that same proposition. All importantly include a conative element related to the believed attitude of the other agent.

The importance of the representation of posture is in its enablement to potentially then manipulate its maintenance or change. From experience, human agents build up beliefs concerning the effects or consequences of the various postures. Perhaps, for example, having a belief about the divisiveness of conflict between two parties being in a third parties favour, may lead to a goal for this to occur. An example of this is Iago recognising the benefits to himself of conflict over love and loyalty between Othello and Desdemona in Shakespeare's "Othello". In other cases, the same belief about conflict may lead to a desire for its resolution; perhaps in a manner which benefits all parties. Automated agents with such beliefs may also generate goals for one posture or another. By representation of the alternatives they have the means of recognition of what exists, as well as a prescription for the requisite alternative changes in attitudes.

From the criticisms of existing notions of cooperation in artificial intelligence, it was argued in chapter 3, that conflict needs acknowledging as a component of cooperative multi-agent interactions. It was also suggested that conflict actually plays an important role in the maintenance and evolution of such cooperation. This chapter will provide the theoretical background to these ideas from social

psychology. First however, some justification is needed for the incorporation of theories and ideas concerning the role and nature of conflict and cooperation in human society, to distributed artificial intelligence (DAI) or human computer interaction (HCI).

5.2 Justification of social theories for computer applications

The volume of research into the structures of multi-agent systems and the nature of multi-agent interactions is currently increasing, especially for distributed processing applications. Important issues for example, relate to the possible characterisations of the nature of organisations of agents - should there be hierarchical structures such as "master/slave" relationships between nodes, what is the distribution of control and decision-making, how are resource limitations to be coped with, what are the alternative structures and means of communication, and so on. Many researchers have considered human societies to be examples of organisations from which potential solutions can be found to these issues, for particular applications and domains; human society can be taken as a paradigm for generating a model of distributed systems.

An example is the work of Fox, who suggests that the designers of distributed systems should draw upon the ideas of other fields "which have considerable experience with their own distributed systems", such as biology or management science (Fox, 1981). He recognises that issues such as motivation, "a module's ability to decide when and what problems to work on" arise, once processing is distributed between separate units in a system. Self-motivation then leads to goal conflicts. He views distributed systems as analogous to human organisations in order to apply the concepts and theories of the management science, organisation theory, to these problems (Fox, 1981). Another example is the work of Gasser, which is quoted in Sridharan's report on the 1986 workshop on DAI as attempting to understand and emulate human production and problem-solving activities, but focussing on "aggregates" rather than single individuals. His aim is a theory of interaction and social organisation for representation in future multi-agent systems (Sridharan, 1987).

However, there is the usual divide between those whose interest is in merely displaying the "competence" of human organisations, over those who wish to also model their methods. The latter are

described by Rosenschein as "the psychological school" (Rosenschein, 1985). An example of this is the work of Doran. His interest is in fact not application oriented, but concerns the use of computational models to learn more about human society, such as the emergence of early human organisation (Doran, 1987b).

Whether aiming at future implementations of computer systems for DAI and HCI applications, and with psychological plausibility or merely displaying behavioural competence, it seems that the analogous study of the nature and properties of human multi-agent organisations, has been considered by others in the field to be an appropriate research strategy. This being the case, examining the potential parallels between theories concerning the relationship of conflict to cooperation in human societies, and these issues in automated systems, has been the methodological approach adopted here. The theories of social conflict considered as appropriate to this research, are to be found in section 5.7. Firstly, the terms conflict and cooperation need precise definition.

5.3 The nature of conflict

In attempting to define conflict as a property of multi-agents, it is firstly assumed that multi-agent conflicts have certain generalisable properties. These properties should characterise each and every conflict, and are therefore entirely discernable from the context.

> ".. there are enlightening similarities between, say, maneuvering in limited war and jockeying in a traffic jam, between deterring the Russians and deterring one's own children, or between the modern balance of terror and the ancient institution of hostages." (Schelling, 1960).

There are a number of such properties which have been compiled from various sources amongst the social psychology literature on this subject. These are elaborated in the rest of this section, following which the ideas are reasserted in terms more appropriate to the nature of this research, and used in the construction of a formal definition in section 5.3.1.

(i) Social conflict requires at least two parties, or two analytically distinct units or entities. Even one party conflict is analysed according to actor versus the environment, or actor versus nature (Mack & Snyder, 1971).

(ii) Mutually exclusive and/or mutually incompatible values and opposed values are inevitable characteristics of conflict (Mack & Snyder, 1971).

(iii) Mutually exclusive, mutually incompatible and opposed values arise from resource limitations. These can be divided into two categories. Firstly, "resource scarcity" occurs when the supply of desired objects or states of affairs is limited, so parties cannot have all they want of anything. Secondly, "position scarcity" describes truths such as that an object cannot occupy two places at the same time, an object cannot simultaneously serve two different functions, a role cannot simultaneously be occupied or performed by two different actors, and different prescribed behaviours cannot be carried out simultaneously (Mack & Snyder, 1971).

(iv) The incompatibility which arises can be either of goals (referred to as conflicts of interests or "ends") and which are described as motivational factors to conflict (McClintock, 1977), or of beliefs. The latter are referred to as conflicts of "means" and are the cognitive factors to conflict (Brehmer, 1977). There may be an interplay between these two factors, such as conflicts of interests deriving from ideological differences or contrasting ideologies developing out of a conflict of interest (Druckman, 1977).

(v) There must be contact between the parties. This does not necessarily have to be "face to face", but involves a "visibility" or awareness of the situation (Mack & Snyder, 1971).

(vi) A conflict relationship always involves the attempt to acquire or exercise power, or the actual exercise or acquisition of power. Power is defined as control over decisions, or the disposition of scarce resources, and the basis of reciprocal influence between or among parties (Sheppard, 1954).

(vii) Conflict cannot exist without action (Kerr, 1954). An action/reaction/action sequence must

embody this pursuit of power.

(viii) Conflict relations "constitute a fundamental social-interaction process". They have important functions and consequences (Dubin, 1957).

> "A conflict process or relation represents a temporary tendency toward disjunction in the interaction flow between parties, but these do not continue to the point where interaction is completely disrupted - the conflict process is subject to its own rules and limits - there is a shift in the governing norms and expectations" (Singer, 1949).

5.3.1 A formal definition of multi-agent conflict

From the ideas concerning social conflict presented in the previous section, multi-agent conflict is formally defined in this section, using the logic which was described in chapter 4.

Firstly, as pointed out in (iv) above, there are two types of conflict - conflict of beliefs and conflict of goals. Point (i) also specifies that there must be more than one agent, and conflict is correspondingly defined as a type of relation between more than one agent and a proposition.

Assumption 2:

$$(\text{CONFLICT } x\ y\ p) = (\text{B-CONFLICT } x\ y\ p) \vee (\text{G-CONFLICT } x\ y\ p)$$

The conflict relation is one between an agent x and another agent y with respect to the proposition p, and is either a conflict of goals or a conflict of beliefs. Each of these is then defined as one of two possible patterns of mental states:

Def 3: $(\text{B-CONFLICT } x\ y\ p) = def$

$$(\text{B-CONFL-I } x\ y\ p) \vee (\text{B- CONFL-M } x\ y\ p)$$

and

Def 4: (G-CONFLICT x y p) = *def*

$$(\text{G-CONFL-I } x\ y\ p) \vee (\text{G- CONFL-M } x\ y\ p)$$

Points (ii), (iii), and (iv) in the previous section, all refer to "mutually exclusive", "mutually opposed" or "mutually incompatible" values. These are terms used in the context of studies in social psychology, and refer to the causes of a conflict. "Mutually exclusive values" in their terms, refer to the desires of two agents to simultaneously possess the same object, or occupy the same space at the same time, for example. It is obviously the case that two agents cannot physically ever occupy the same space at the same time, and that this expresses a notion of mutual exclusivity BUT, the beliefs about or desires for these states are not mutually exclusive. They can and do co-exist. They are not logically incompatible.

The definition of conflict is a description; it describes such a state of affairs. It merely describes the mental states of an agent regarding her beliefs about her own and another's beliefs or goals with respect to the proposition in question. The primary (but not sole) condition of conflict is therefore that the other agent is believed to have a belief or goal which is in opposition to her own. For example, (BEL x p) and (BEL x (BEL y ~p). These beliefs comprise the element of "visibility" or "awareness of the situation", as suggested in point (v) in the previous section.

Def 5: (B-CONFI-I x y p) = *def* $(\text{BEL } x\ p) \wedge (\text{BEL } x\ (\text{BEL } y\ \sim p))$

$$\wedge\ (\text{P-R-GOAL } x\ (\text{BEL } y\ p)\ q)$$

This says that x has a belief p, and x believes that y believes not p. x also has a P-R-GOAL that the other eventually change their belief, only to be abandoned if this is achieved, becomes impossible to achieve or it is no longer the case that q, q being a reason for x having this goal.

Def 6: (G-CONFI-I x y p) = def $(\text{GOAL } x\ \Diamond p) \wedge (\text{BEL } x\ (\text{GOAL } y\ \Diamond \sim p))$

$$\wedge\ (\text{P-R-GOAL } x\ (\text{GOAL } y\ \Diamond p)\ q)$$

This says that x has a goal for p to be true from now into the future, and believes y to have the goal eventually not p. She also has a persistent goal for y to change her mind and therefore take on the goal

eventually p, relative to q.

The P-R-GOAL is essential to the characterisation of conflict. It indicates a commitment to action to change the other agents mind. This reflects points (vi) and (vii) above. If the P-R-GOAL is abandoned this may be because it is achieved, but it may also be abandoned because q is no longer true, or the agent believes it has become impossible to achieve. Whatever the reason, if the P-R-GOAL is abandoned it is irrelevant whether x still believes y to believe or want the opposite to her; the situation is no longer conflict. A particular incident of conflict is over when no agent in the system has a goal for further change with respect to the issue in question, whether the goal has in fact been satisfied or not. What is achieved from such a resolution of the conflict is a renewed, temporary stability of the multi-agent system; there is currently no goal for change. This ties in with point (viii) concerning the functions and consequences of conflict.

It is important to note that conflict is defined here as being subjective. In other words, x only needs to believe that y believes a proposition which is the opposite to her own belief or goal with respect to that same proposition, and have a committed goal to change this. The reality of y's belief in relation to p is irrelevant.

The alternative patterns of beliefs and goals defining belief and goal conflicts in defs 3 and 4, are as follows:

Def 7: (B-CONFL-M x y p) = def (BMB x y ((BEL x p) $\wedge$ (BEL y ~p)))

$\wedge$ ((P-R-GOAL x (BEL y p) q) $\vee$ (P-R-GOAL y (BEL x ~p) q))

This says that it is mutually believed between x and y that they have a difference in belief related to p. Either one or both of them have a P-R-GOAL to eventually change the others belief, only to be abandoned if this is achieved, becomes impossible to achieve or it is no longer the case that q.

Def 8: (G-CONFL-M x y p) = def (BMB x y ((GOAL x ◊p) ∧ (GOAL y ◊~p)))

∧ ((P-R-GOAL x (GOAL y ◊p) q) ∨ (P-R-GOAL y (GOAL x ◊~p) q))

This says that it is mutually believed between x and y that they have a difference in goal, such that one has a goal for p whilst the other has a goal for not p. Either one or both of them have a P-R-GOAL to eventually change the others goal, only to be abandoned if this is achieved, becomes impossible to achieve or it is no longer the case that q.

N.B. p may represent events in the physical world performed by agents, for example (DONE y a) or (DONE x a), or changes in another's mental states, such as (BEL y r).

The definition of multi-agent conflict therefore is:
conflict of goals or conflict of beliefs exist between one agent and another, when the agents' beliefs or goals with respect to the same proposition are believed by the one agent to be in opposition, and this agent also has a persistent goal to change the other's belief or goal. Alternatively, there may be a mutual belief about the difference in belief or goal between the participating agents, in which case, conflict exists if either or both also has a persistent goal to change the other's belief or goal.

There are some types of human conflicts which do not apparently conform to this definition. In these, the goal is not to persuade the other to change their belief or goal; actions are performed as a means of maintaining the conflict. Perhaps the participants enjoy the "banter", or it may serve purposes of assertion of values for the benefit of a third party. Such dialogues are better reasoned about from the point of view of the participants as cooperative, and therefore having a common goal to maintain disagreement. The goals upon which one or both agents are acting, are not to genuinely get the other to change their mind. Even so, if during such an interaction one agent capitulates on a particular issue, then an incident of conflict is over, even if another one is instantly started in order to satisfy the goal of

maintaining disagreement. Recognising another's goal to maintain disagreement if one does not have that goal oneself, is itself another conflict if there is also a committed goal to get the other to stop the row.

5.4 The nature of cooperation

A large part of the discussion surrounding the nature of cooperation has already taken place in chapter 3. In summary, previous work in artificial intelligence has considered the following to be characteristic of cooperative interaction:

(i) a common goal exists between cooperating multi-agents,
(ii) at least a component of possessing the goal is an awareness that another agent has it as a goal,
(iii) recognition of another's goal is sufficient to adopt that goal as one's own, assuming no contradictory one already in existence, and
(iv) if the attainment of the common goal requires the attainment of sub-goals, these are adopted and shared out between the interacting parties, on the same basis as above.

It was claimed in chapter 3, that this characterisation is one of benevolence, not cooperation; that truly autonomous agents adopt other's goals not merely because they are made aware of their existence, but also based upon their own preferences. Cooperation therefore can be achieved even in contexts involving conflict, by the creation of circumstances (using dialogue) which can conform to another's preferences. This is in contrast to the majority of existing work in this area which unrealistically assumes the non-existence of differences or conflicts, and secondly ignores its potential benefits to cooperative multi-agent environments. Cooperation by an agent x with respect to another autonomous agent y and some proposition p, alternatively has the following characteristics:

(i) cooperation requires a common goal between agents. Recognition of another's goal may lead to this situation of a common goal. This would be dependent upon the preferences of the receiving agent.

This allows another's goal to be cooperatively adopted even if there is a contradictory one already in existence. The conditions for this are elaborated in chapter 6 as Assumption 5. Alternatively, it may be mutually believed that there is already a common goal in existence,

(ii) an essential component of x's cooperation is her awareness that her goal is relative to y's having the goal. In this way cooperation is more than mere accidental coordination, as described in section 3.2.2.1, but incorporates an element of helpfulness,

(iii) cooperation can be partially summarised as a recognition or belief about another's goal and a personal preference for this goal state to be achieved, and as a consequence of both of these conditions, there is a commitment to achieving the common goal, and

(iv) if attainment of the common goal requires the attainment of sub-goals, these are adopted on the same basis as above.

Def 9:

(COOPERATION x y p) = *def*

$$(\text{COOP-I } x\ y\ p) \vee (\text{COOP-M } x\ y\ p)$$

Cooperation is defined as a relation between an agent x, with respect to some other agent y and a proposition p. The definition comprises two alternative patterns of mental states. Here again, p can represent events in the physical world performed by agents, such as (DONE y a) or (DONE x a). Alternatively it may represent changes in another's mental states , such as (BEL x r).

The first possible pattern of mental states reflecting x's cooperation with respect to y and the proposition p relates to the common goal being as a result of x recognising y's goal:

Def 10: $(\text{COOP-I } x\ y\ p) = \text{def} \quad (\text{BEL } x\ (\text{GOAL } y\ \Diamond p)) \wedge (\text{PREFER } x\ p\ \sim p) \supset (\text{P-R-GOAL } x\ p\ (\text{GOAL } y\ \Diamond p))$

This says that cooperation for x with respect to y and p is the recognition of y's goal that p be eventually true and x preferring p to ~p, resulting in the generation of a persistent goal for p, relative to y's possession of the goal. In other words as long as p is not achieved, x believes p possible to achieve, and x believes y to have the goal ◊p, then x is committed to the goal generated from her own

preferences on recognition of it as another's.

The second possible pattern of mental states reflecting x's cooperation with respect to y and the proposition p relates to there being a mutual belief between x and y that they have a common goal:

Def 11: (COOP-M x y p) = def (BMB x y ((GOAL x ◊p) ∧ (GOAL y ◊p)))

⊃ ((P-R-GOAL x p (GOAL y ◊p)) ∨ (P-R-GOAL y p (GOAL x ◊p)))

This says that if there is a mutual belief between x and y that they both possess the goal that eventually p, then being cooperative means either one or both have persistent goals relative to the existence of the other's goal. Preferences are unnecessary to this definition. Both agents autonomously possess the goal, and believe this of each other. All that is required is to distinguish between cooperation as a situation comprising a common goal, and accidental coordination. This is provided by their commitment to the goal being relative to the other's possession of it.

The definition of cooperation between multi-agents therefore is: cooperation exists between one agent and another with respect to some proposition when one agent recognises the other's goal that this proposition be realised and, as a consequence of also a personal preference for this goal state to be achieved, is committed to achieving it as a common goal, relative to the other agent having it as a goal. Alternatively, cooperation exists between one agent and another with respect to some proposition if one or both is committed to achieving the goal for this proposition to be realised relative to the other agent having it as a goal, as a consequence of it being mutually believed that they have a common goal.

5.5 The nature of indifference

The term "indifference" implies a lack of caring. In this case the lack of caring is about another agent's attitude towards the proposition in question. This means the conative element in the postural definition is that there must be a lack of goal in relation to the other agent's attitude. For example, x may believe that she believes p and y believes not p: (BEL x p) and (BEL x (BEL y ~p)). If x has no goal to change this state of affairs, then she is indifferent with respect to y and p. Likewise, if x recognises that y has a goal p, then even if x also happens to have the goal p, unless she commits herself to this because y has it, and therefore she has a goal which in some way relates to y's possession of the goal, she is indifferent with respect to y and p.

Indifference is defined as follows:

Def 12:

(INDIFFERENCE x y p) = def

$$[\sim(\text{GOAL } x\ \Diamond(\text{BEL } y\ p)) \wedge \sim(\text{GOAL } x\ \Diamond(\text{GOAL } y\ p))\ ]$$

$$\vee\ \sim(\text{P-R-GOAL } x\ p\ \Diamond(\text{GOAL } y\ p))$$

This says that x being indifferent with respect to y and p is defined as x not having a goal for y to believe or have a goal for p, or herself not having a goal for p which is relative to y's goal for p.

5.6 Mixed postures in the multi-agent system

It is feasible that a context could exist where one agent x believes themselves to be in conflict with another y, and therefore has a persistent goal to change y's mind in relation to p, and the other agent y simultaneously believes herself to be cooperative with respect to x and p. She adopts x's goal as her own. Alternatively she may be indifferent to x and p. Alternatively again, there may be no postural relation between y and x and p; y may have no attitudes with respect to p, or have no beliefs concerning x's relations to p. An example of such a context where one agent is in conflict with another who is unaware of the situation and has no posture with respect it, is provided by Iago and Othello, as described in chapter 7. Iago wants revenge on Othello; he wants to cause him harm. He believes that Othello does not want this, but generates a goal for Othello to believe his wife is unfaithful to him, which will inevitably cause great personal distress. Othello initially has no beliefs regarding Iago's beliefs about his wife Desdemona's, fidelity. The postural relation is only Iago's with respect to Othello and the proposition regarding Desdemona's fidelity. As a result of the conversation however, Othello recognises not only that Iago has a belief about Desdemona's fidelity, but has a goal for Othello to adopt this. Othello's postural relation becomes one of cooperation when he adopts Iago's goal to believe Desdemona unfaithful.

5.7 The role of conflict in multi-agent systems

The fundamental challenge which this research poses to previous research on cooperative interaction in artificial intelligence, relates to the notion of cooperation within which multi-agent systems are designed to operate. It was pointed out in chapter 3, that conflict has generally been ignored in previous work; cooperative multi-agent systems adopt each other's goals benevolently, and if conflicts are acknowledged, they are avoided. The value of Rosenschein's work (Rosenschein, 1985, Rosenschein & Genesereth, 1985), has been to point out the ubiquity of conflict in everyday cooperative interactions and the necessity for practical real-world applications involving cooperative multi-agents to be flexible enough to be able to reason about action in the light of conflict.

This research comprises a theory of multi-agent interaction which in agreeing with Rosenschein, goes somewhat further in adopting a view of conflict and cooperation derived from social psychology. Conflict has positive and important functions in the maintenance and evolution of cooperation.

A justification of the use of theories regarding social conflict between humans as pertinent to studies concerning automated multi-agents was given in section 5.2. Some suggested positive roles of conflict to the maintenance and evolution of cooperative multi-agent systems are therefore described in the sections which follow.

5.7.1 The positive functions of conflict

> "Conflict is a pervasive and inevitable aspect of life. Its pervasiveness suggests that conflict is not necessarily destructive or lacking in pleasure. Conflict has many positive functions. It prevents stagnation, it stimulates interest and curiosity, it is the medium through which problems can be aired and solutions arrived at; it is the root of personal and social change. Moreover, conflict is often part of the process of testing and assessing oneself and, as such, may be highly enjoyable as one experiences the pleasure of the full and active use of one's capabilities" (Deutsch, 1971).

In the above quote, Deutsch stresses especially the personal gains from conflict. He appeals to the

fact that individuals actively seek out conflict in competitive games, at the theatre, in novels, in intimate encounters, at work and so on, to suggest that such conflict is beneficial to those individuals. It is not a "villain", "... the cause of psychopathology, social disorder, war" (Deutsch, 1971). Psychological utopia is not a conflict-free existence.

A broader and more complete description of the positive functions of social conflict is provided by Coser (Coser, 1956). He derives his ideas from the classical work of Georg Simmel (1858 - 1918). It is interesting to note that whilst commenting that early American sociologists such as Charles Cooley, Edward Ross, William Sumner and others, also regarded conflict as constructive and functional, Coser expresses concern that his contemporary sociologists viewed conflict as dysfunctional. He refers to the work of Talcott Parsons, Elton Mayo, George Lundberg as focussing on maladjustment and tensions, and these as interference to concensus. However, by the time Coser's second book appeared, a decade later (Coser, 1967), his views were being widely endorsed. Nowadays he is rarely quoted in literature on social conflict; his ideas are assumed. The following is a list of Coser's views on the positive functions of conflict :

Conflict exerts pressure for innovation and creativity in social systems (Coser, 1957). It thereby enables a shift in the governing norms and expectations; it revitalises existent norms or contributes to the emergence of new ones. This is important to the maintenance and evolution of social systems because such systems are dynamic; survival in constantly changing conditions requires regular reevaluation and rebalancing (Coser, 1956). This principle is also applied in biological theories of social adaptation, evolution and survival. Male stags fighting each year for example, is a regular conflict ensuring that either the existing power structure is reestablished by the same male remaining dominant, or else he is overthrown and a new order established.

> "... social conflict is a mechanism for adjustment of norms adequate to new conditions. A flexible society benefits from conflict because such behaviour, by helping to create and modify norms, assures its continuance under changed conditions. Such mechanisms for readjustment of norms is hardly available to rigid systems: by suppressing conflict, the latter smother a useful warning signal, thereby maximising the danger of catastrophic breakdown." (Coser, 1956).

Internal conflicts therefore ascertain the relative strengths of antagonistic interests within a structure, and thus are a mechanism for the adjustment or maintenance of the balance of power. The resolution of such conflicts establishes a new equilibrium. Coser (1957) quotes a natural scientist Kaemfert, who in an article in the New York Times in July 1952, put forward similar views with respect to the functions of earthquakes. Earthquakes were suggested as being the earth's way of maintaining equilibrium, an adjustment enabling the crust to yield to stresses which may reorganise or redistribute materials.

Intergroup conflicts set group boundaries by strengthening group cohesiveness and separateness. Conflicts involving associations or coalitions provide a bond between the members, uniting and removing social isolation. A social structure in which a multiplicity of conflicts can exist with associations whose "diverse purposes crisscross each other" prevents "alliances along one major line of cleavage" (Coser, 1956).

Conflict also reduces tension and permits maintenance of social interaction under stress. It clarifies objectives. It allows agreement rather than subordination.

Coser concludes:

> "conflict tends to be dysfunctional for a social structure in which there is no or insufficient toleration and institutionalization of conflict ... What threatens the equilibrium of such a structure is not conflict as such, but the rigidity itself which permits hostilities to accumulate and to be channeled along one major line of cleavage once they break out in conflict" (Coser, 1956).

5.7.2 Conflict, autonomy, and the evolution of cooperation

Axelrod has examined the question of the evolution of social cooperation, "in a world of egoists without central authority" (Axelrod, 1984). In other words, if each member of a multi-agent system has their own self-motivated goals, and there is no central authority exerting insistencies concerning benevolence towards others, can cooperation emerge? According to Axelrod, the answer is "yes". This, of course, is very relevant to the proposal here that automated multi-agents need to be truly autonomous

with control over their mental states, and to experience conflicts as a part of being cooperative agents operating in the real world.

Axelrod's theory is based upon investigations into individuals pursuing their own self-interest without any enforced cooperation. He points out that self-interest does not imply complete abandonment of concerns for others. He offers interactions between a brother and sister, or friendly nations, as examples.

> "The assumption of self-interest is really just an assumption that concern for others does not completely solve the problem of when to cooperate with them and when not to." (Axelrod, 1984).

The classic game, the Prisoner's Dilemma was Axelrod's chosen means of analysing the problem. Each player has something to gain both from cooperating and being exclusively self-motivated. The fundamental issue for the Axelrod's Cooperation Theory, is the number of times the game is played. Two egoists playing the game only once will choose their dominant move of defection and disclosing information to the police, although each therefore gets a worse outcome than if they had both cooperated. There is also no incentive to cooperate if the game is played more than once, but a known finite number of times. BUT - this is not the case, if the players will interact an indefinite number of times.

> "The evolution of cooperation requires that individuals have a sufficiently large chance to meet again so that they have a stake in their future interaction. " (Axelrod, 1984).

This argument was referred to in section 3.2.4 whilst discussing the employment of mendacity as a strategy in conversation. It was suggested that the beliefs agents hold regarding potential future effects of their actions, both good and bad, will affect the determination of their preferences, and therefore also their goals. The issue surfaced again in section 3.2.5.1, whilst discussing Richelson's metagame analysis of the nuclear/conventional warfare bluff between the United States and Soviet Union (Richelson, 1979). His payoff analysis related only to the immediate gains and losses of the various alternative strategies. Agents engaged in the kinds of cooperative tasks for which this theoretical research is envisaged as being appropriate, should have the potential of future interactions included in

their reasoning as to appropriate dialogue action. The immediate future is catered for in the inclusion of expectations as to subsequent action as a result of the intended dialogue action if employing a strategic rationality. More general expectations are embodied in the determination of preferences as described in chapter 2.

Axelrod's research included the instigation of a computer tournament for the Prisoner's Dilemma Game. The programs were designed to select moves according to a history of the game so far, in a variable number of games. Each entrant was programmed to use a different strategy. The outright winner each time was the simplest; it was a program written by A. Rapaport comprising a TIT FOR TAT strategy. This starts with a cooperative choice and after that simply does what the other player did on the previous move.

> "What accounts for TIT FOR TAT's robust success is its combination of being nice, retaliatory, forgiving, and clear. Its niceness prevents it from getting into unnecessary trouble. Its retaliation discourages the other side from persisting whenever defection is tried. Its forgiveness helps restore mutual cooperation. And its clarity makes it intelligible to the other player, thereby eliciting long-term cooperation." (Axelrod, 1984).

Axelrod's research also includes the investigation of cooperation in non-human organisms, for which he refers to the theories of biologists Maynard-Smith (Maynard-Smith & Price, 1973) and Dawkins (1976). He suggests that in accordance with the predictions of his theory, almost all clear cases of altruism and most observed cooperation between animals, occur in the context of high-relatedness, usually between immediate family members. This is genetical kinship theory which takes " a gene's-eye view of natural selection (Dawkins, 1976)" (Axelrod, 1984). The evolution of cooperative behaviour between organisms where relatedness is low, such as in examples of symbiosis, is explained with a theory of reciprocity:

> "When the probability of two individuals meeting each other again is sufficiently high, cooperation based on reciprocity can thrive and be evolutionarily stable in a population with no relatedness at all" (Axelrod, 1984).

Further examples for this are taken from trench warfare in World War 1, and the tacit cooperation which evolved from one side being the first to cooperate and then an arrangement of reciprocation continuing (Axelrod, 1984).

5.8 Conclusions

It is acknowledged that the model of agents incorporated in this research is not a model of human agents. The nature of the model and its limitations have already been discussed in chapters 2 and 4 . The definitions of conflict and cooperation necessarily therefore also suffer in being "less than human". However, the parallels which are being made from examining human systems, concerning the role of conflict in the maintenance and evolution of cooperative multi-agent interaction seem valid, regardless of this distinction. Rapaport's computer tournament and Axelrod's examples from biological systems (which presumably also have limited cognitive abilities), offer demonstrations of general laws or principles of multi-agent interaction, human or non-human, at work.

To summarise how these principles are integrated into the framework for future computer models of cooperative dialogue as proposed by this research:

Interacting agents have representations of conflict, cooperation and indifference as three alternative postures which may exist between them, with respect to any proposition. These comprise patterns of mental states. Existing postures can thereby be recognised, and in association with other known properties of interacting agents as described in chapter 3, other postures can be desired with commitment. One of these properties concerns dialogue as strategic interaction; agents adopting a strategic rationality with which to attain these postures, and correspondingly either individual or collective goals. The nature of the representations is such that both conflict and cooperation are understood as states from which goals are achieved or dropped, and there may therefore no longer be a commitment to action to change the state of the system. In the case of cooperation, the goal achieved or dropped is a common goal, thus satisfying all parties; with conflict the lack of commitment to action will either have maintained the

status quo, or established some new relation, but at this point in time, no member of the system still wants to challenge it. This stability is temporary given the dynamic nature of social systems in real environments. However, if each time conflicts arise they are expressed and resolved, as opposed to being suppressed or avoided, the system as a whole can survive and be flexible in a changing and unpredictable environment.

It was explained in section 5.7.2 that expectations of future interaction are essential to the manner in which conflict resolution and therefore this evolution of multi-agent systems can occur. Beliefs relevant to the determination of preferences, associated especially with expressions such as mendacity and concealment, provide an element of "morality" in the determination of strategy. These beliefs come from within and interact with the immediate task-related goals. They allow autonomous and pragmatically determined reasoning as to appropriate action in the particular circumstances. **This offers flexibility of action and the evolution of true cooperation to the advantage of the system as a whole, where imposed benevolence does not.**

> "Conflict, of some sort, is the life of society, and progress emerges from a struggle in which individual, class, or institution seeks to realize its own idea of good" (Cooley, 1909).

CHAPTER 6 : Control of Information, Strategic Interaction and the Autonomy of Agents

6.1 Introduction

Pragmatic theories consider dialogue as comprising utterances. For this framework, utterances have been described in chapter 1 as speech actions performed by agents to convey the information that the speaker is in a particular mental state, which relates to the fact that the speaker has a goal to induce a particular mental state in the hearer. Dialogue is therefore intentional communication; the agent is not merely behaving in some way without any thought of what is being conveyed to others. An example of such unintentional communication might be an agent conveying the information that she is nervous or hot by sweating (Allwood 1976).

Generating utterances may be a form of intentional communication, but between human agents, there are no guarantees with respect to the attainment of the intended or goal state. Just because one agent acts upon a goal to induce a particular belief state in another, and the receiving agent recognises this, she still may not actually take on that belief. The acting agent only has partial control over the effects of her dialogue actions because the receiving agent has total control over the mental states she acquires; she is autonomous. The speaker's autonomy on the other hand, is reflected in her ability to reveal or conceal, and truthfully or deceitfully represent her mental states. She has the control over the divulgance of the true nature of her mental states.

Autonomous agents share control over the flow of information between them in dialogue.

In this framework, cooperative dialogue is generated and interpreted according to an understanding of such control issues. Each agent in the multi-agent system, whether human or automated, is believed by themselves and others to have the potential to alter the mental states of others, but taking into account the control that each autonomous agent also has over what they reveal and what they acquire. Strategic interaction acknowledges the interdependence between agents when reasoning about dialogue action and

its effects.

These issues of control of information, strategic interaction and the autonomy of agents, were raised in chapter 3. They were introduced in the context of the theory of multi-agent interaction proposed in this thesis as the basis for reasoning about cooperative dialogue action. The role of this chapter is to describe these ideas more fully, and to express them formally in the language described in chapter 4. The resulting axioms are formal statements regarding the nature of belief and goal adoption in strategic dialogue, certain properties attributable to speech acts regarding openness and truthfulness, and assumptions concerning the use of these. In addition to the representations of the nature of the three postures as defined in the previous chapter, these are more of the proposed extensions to the formal model of agents outlined in chapter 4, with which agents can reason about cooperative dialogue.

The term "information" is very briefly clarified in section 6.2. Section 6.3. deals with autonomy and control of information with respect to the acquisition of mental states in dialogue, and section 6.4 deals with these same issues, but with respect to the revealing of mental states in dialogue. Comparisons are made with the approaches of Cohen and Levesque (1987b) and Perrault (1987). This is to point out the suggested advantages of this framework's notion of autonomy as a means of achieving multi-agent cooperation by negotiation, and thus acknowledging the positive role of conflict. The contrast is made in section 6.3.1.1 with respect to Cohen and Levesque's (1987b) and Perrault's (1987) conditions for belief and goal adoption in dialogue, which incorporate notions of sincerity and helpfulness. These are embodied within a characterisation of an example speech act, which is compared with another such characterisation incorporating instead, elements of this framework's theory of multi-agent cooperation, in section 6.5. The alternative treatments of insincere or non-serious acts such as lies and irony, are also contrasted towards the end of section 6.3.1.2.

6.2 Information

The term information is described by Dretske (1981) as follows:

"Roughly speaking, information is that commodity capable of yielding knowledge, and what

information a signal carries is what we learn from it."

Therefore, if y performs an utterance from which y wants x to eventually believe p, and x correctly recognises this goal, then this is information conveyed. x did not previously believe that y wanted him to believe p; x has learnt from the dialogue. If he incorrectly recognises the goal however, according to Dretske he has not been informed. He cannot learn that y has a goal which y does not have. Agents cannot know what is false, and information yields knowledge. However, the utterance may not be entirely uninformative; he may acquire other correct beliefs or knowledge, such as that y was lying, for example.

Allwood uses the term "information" slightly differently:

> "Information will be used as an abstract term for any object that could be apprehended with some degree of alertness by a conscious agent ... Further, an object is informative iff either the object itself, or some other object connected with it, *is apprehended by the agent.*" (Allwood 1976).

The term "object" includes abstract entities such as numbers, colours, as well as those which are more concrete, and apprehension refers simply to the conscious attention of the receiving agent.

According to this thesis' framework, information is also considered that which can be apprehended by an agent. What is acquired as a result of this apprehension, is a mental state. The agent has a mental state which she previously did not have; she has learned. Dretske's requirement that the learning only be considered valid if what is learned is true and therefore the agent's gain be in terms of knowledge, is rejected here. In this framework, the yield is a mental state acquired, whether this be knowledge or belief. Sperber and Wilson also use the term "information" not only as facts but "dubious and false assumptions presented as factual" (Sperber & Wilson, 1986). Again in accordance with Allwood, individuals are considered as able to convey information both intentionally and unintentionally. However, since the focus is on generation of dialogue action and its interpretation via recognition of communicative goals, further discussion is restricted to intentional communication. It is also acknowledged that beliefs and goals can be adopted on occasion, at an unconscious level, such as with subliminal perception (Allwood 1976) in subliminal advertising. However, the adopting/acquiring of a

belief is only considered here conditionally upon conscious apprehension of the conveyance of information from another agent.

6.3 Control over information acquired in dialogue

An agent performing a dialogue action as a result of having a goal to effect a particular change to the mental states of other agents in the multi-agent system, may or may not achieve this objective. Allwood gives an example of A intending B to believe he is an Arab by wearing a burnoose. However, B may believe he is on his way to a fancy dress ball instead (Allwood 1976). Alternatively, the goal can be recognised but not adopted. For example:

> "I can successfully assert that it is cold in here without convincing you of that fact" (Perrault 1987).

In other words, agents in multi-agent systems do not have total control over the effects of their actions. It is not enough to select appropriate speech acts on the basis of goals alone; the receiving agent(s) also have partial control over the effects of speaker's actions. Decisions about the generation of appropriate dialogue action should be made in the light of this understanding that agents are autonomous over their own mental states. Representation is also therefore required of a property of interacting agents :

recognition of another agent's goal to induce a particular mental state results in the hearer believing that eventually she will either have adopted that mental state or not adopted it.

Assumption 3:

$(BEL\ x\ (GOAL\ y\ \Diamond q)) \supset (BEL\ x\ (\Diamond q \vee \Diamond \sim q))$

For example:

$$(BEL\ x\ (GOAL\ y\ \Diamond(BEL\ x\ p))) \supset (BEL\ x\ (\Diamond(BEL\ x\ p) \vee \Diamond \sim(BEL\ x\ p)))$$

p represents the proposition that it is cold. If x recognises y's goal for her to eventually believe that it is cold, then x believes she will either eventually believe that it is cold or eventually not have a belief about it. Similarly:

$$(BEL\ x\ (GOAL\ y\ \Diamond(DONE\ x\ a))) \supset (BEL\ x\ (\Diamond(DONE\ x\ a) \vee \Diamond \sim(DONE\ x\ a)))$$

a represents the action of washing up. If x believes y has a goal that x eventually have done some washing up, x believes she will either eventually have done some washing up, or not.

In order to be more useful, an axiom is required which states that: recognition of another agent's goal to induce a particular mental state results in the adoption of that goal, and therefore potentially the desired mental state being actualised, if certain conditions hold. What are these conditions? This question is answered in the following sections. Dialogue as strategic interaction is assumed.

6.3.1 Conditions for belief and goal adoption in strategic interaction

6.3.1.1 An introductory discussion of the issues - a comparison with Cohen and Levesque's approach

The question of the general conditions under which agents will adopt a belief or goal, was originally discussed in section 2.2.2.1. The conditions under which this occurs in the context of dialogue, and therefore recognition of another's goal that the belief or goal be adopted, incorporates also other aspects of multi-agent interaction, which were briefly discussed in section 3.2.4. In section 6.3.1.2, these ideas will be reiterated and then expressed as formal statements. The rest of this section comprises a discussion of the issues which are relevant to belief and goal adoption for a framework which focusses on negotiation and conflict resolution. It is in the form of a comparison with the approaches of Cohen and Levesque (1987a, 1987b) and Perrault (1987).

Firstly there is the issue of the nature of cooperation, and communication as cooperative interaction.

Cohen and Levesque's description of cooperative interacting agents is of agents benevolently adopting other's desired mental states in dialogue. This means that recognition of another agent's goal to induce a belief in oneself for example, is enough justification for adopting that goal as one's own, and consequently adopting the belief. The only negative condition is the prior existence of a contradictory belief. In Cohen and Levesque's and Perrault's frameworks, as well as this thesis' strategic framework, agents cannot believe p and not p at the same time. There is consistency of beliefs. This means that only if the original belief is dropped, can agents successfully induce in others, beliefs which are contradictory to existing beliefs. But, can an existing belief or goal be dropped as a result of a communicative action? This is an important issue in a strategic framework, because if the adoption of a new belief or goal is always conditional upon it being consistent with existing and unchangeable ones, then there can never be the opportunity to "change someone's mind" ; no amount of discussion will result in an agreement between x and y with regard to p if x already believes or has a goal p, and y has an existing belief or goal, not p.

As described in chapter 1, Perrault explicitly states that existing beliefs persist, whilst acknowledging this as a simplification for the purposes of his own research. There are no conditions in his framework, whereby new beliefs can be adopted which are in contradiction with those in existence. Cohen and Levesque however, do suggest that beliefs can become false, and allow goals to be dropped (Cohen and Levesque, 1987a, 1987b). Their P-R-GOALS for example, are relativised goals which can be dropped for three reasons, one of which is the mental state comprising the reason for their existence becoming false. The question is whether this can occur via dialogue, and under what conditions. Firstly, can an agent drop an existing goal in order to then adopt another contradictory one, as the result of a communication? Communicating agents are described by Cohen and Levesque as cooperative, defined as HELPFUL. They describe HELPFUL agents as follows:

> "An agent is HELPFUL to another if he adopts as his own persistent goal the other agent's goal that he eventually do something (provided that potential goal does not conflict with his own)" (Cohen and Levesque, 1987b).

In the formal definition of HELPFUL, a conflicting goal is shown to be one that the agent believes to

be forever false:

C&L Def 12: (HELPFUL x y) = def

$\forall a$ (BEL x (GOAL y $\Diamond$(DONE x a))) $\wedge$ ~(GOAL x $\Box$~(DONE x a)) $\supset$

[P-R-GOAL x (DONE x a) (GOAL y $\Diamond$ (DONE x a))]

x is defined as a helpful agent with respect to y. x will generate persistent goals to do anything y wants him to have done, as long as x having done the act in question is not desired by x to be forever false or impossible. Commitment to the persistent goal is relative to the existence of y's goal.

The following example is an attempt to show that Cohen and Levesque's helpful agents also have persistent attitudes, which prevent the possibility of negotiated agreements: Suppose that an agent x has a goal to eventually have done a: (GOAL x $\Diamond$(DONE x a)), and also that this goal is inconsistent with some other goal, for example for x to eventually have done b:

(GOAL x $\Diamond$(DONE x a)) $\supset$ (GOAL x ~$\Diamond$(DONE x b)). However, x recognises y's goal for her to eventually have done b: (BEL x (GOAL y $\Diamond$(DONE x b))). According to Cohen and Levesque's definition of HELPFUL, the only condition whereby she would <u>not</u> take on y's goal to eventually have done b as a committed goal, is if a goal for her to have done b is forever false, or impossible: (GOAL x $\Box$~(DONE x b)). Is this the case here? Does her existing goal to have done a imply that a goal to have done b is impossible? Are these goals therefore conflicting? A proposition is forever true iff it is not eventually not true : $\Box$ p $\equiv$ ~$\Diamond$~p (given in section 4.4.3.1) . Correspondingly, it must be forever false iff it is not eventually true : $\Box$ ~p $\equiv$ ~$\Diamond$p. Therefore :

(GOAL x $\Box$~(DONE x b)) $\equiv$ (GOAL x ~$\Diamond$(DONE x b)), which shows that agent x cannot be helpful, and take on y's recognised goal for her to have done b, because her existing goal that she have done a implies that her having done b is forever false. If it were not the case that agent x having a goal to have done b be <u>forever</u> false, then the definition of HELPFUL might have been satisfied in the example above, and x could have adopted the goal to have done b. In fact, agents would then helpfully take on <u>any</u> goal that they believe other agents have communicated to them, unless this is previously

desired to be forever false. This seems an improbably strong notion of benevolence. But, such goals being believed forever false and in conjunction with the characterisation of communicating agents as HELPFUL, means that Cohen and Levesque's agents are unable to change their goals via dialogue; contradictions to existing goals are forever false. This then is equivalent to Perrault's persistence theory where beliefs and goals cannot be revised in dialogue. A lack of disagreement between agents must therefore be assumed by both Cohen and Levesque and Perrault, in order for any cooperative interaction.

The only condition according to which Cohen and Levesque's helpful agents might not take on another's communicated goal, even if they have no existing contradictory one, is if they believe the speaker to be insincere. This sincerity issue is also an important one in the context of strategic interaction. According to Cohen and Levesque (1987b), an agent is sincere with respect to another agent and some proposition p if whenever x has chosen to do something bringing it about that y believes p, x has chosen to bring it about that y knows p:

C&L Def 13: $(\text{SINCERE } x\ y\ p) = \text{def}\quad \forall e\ (\text{GOAL } x\ (\text{HAPPENS } x\ e;(\text{BEL } y\ p)?)) \supset (\text{GOAL } x\ (\text{HAPPENS } x\ e;\ (\text{KNOW } y\ p)?))$

> "x would be insincere to y about p if x wants y to believe something that x wants to be false." (Cohen & Levesque, 1987b).

In contrast to this definition of Cohen and Levesque's, this thesis' framework considers truthfulness only in relation to the agent's belief, rather than the world. It is considered too strong a condition to require sincerity to relate to truth in the real world. In other words, as long as an agent <u>believes</u> a proposition to be true, then an utterance performed by that agent which is an expression of this belief is truthful. This work also does not consider truthfulness to be a property of agents. At best agents can have a disposition or predilection towards truthfulness, but it is again too strong to say that an agent is truthful per se. Therefore veracious, mendacious, concealing and revealing are defined in section 6.4 as properties of acts. They are types of action expressions; types of utterances. In this way, (VERACIOUS e (BEL x p)) for example says that the utterance e is veracious with respect to x's belief p, and at the same time, (MENDACIOUS f (BEL x p)) can also be true. This says that

utterance f is mendacious with respect to x's belief p.

Cohen and Levesque's (1987b) definition of a request encompasses their ideas concerning goal adoption between cooperative agents. It is a complex action expression performed by the speaker to bring about a certain state of affairs:

$$q = \text{def } \sim[\text{BEFORE } [\sim(\text{COMPETENT } y \text{ (DONE } y\ a)) \vee (\text{BEL } y\ \Box\sim(\text{DONE } y\ a)) \vee (\text{BEL } y\ \sim(\text{GOAL } x\ \Diamond(\text{DONE } y\ a))] \sim(\text{GOAL } y \text{ (LATER (DONE } y\ a)))]$$

The hearer only drops the adopted goal because it is achieved.

C&L: Def 14: $\{CA\ x\ e\ p\ q\ r\} = \text{def}$

$$(\text{P-R-GOAL } x \text{ (DONE } x \text{ [GOAL } x \text{ (HAPPENS } x\ e;\ (p \wedge q)?)]?\ ;\ e;p?)\ r);e$$

CA is a complex action expression such that e is a committed attempt by x to achieve p. e is done when x has a persistent goal to do e to achieve the intended effects p. q is another effect of e which is not immediately intended but is also a goal of x's. From the definition of P-R-GOAL, if x does not achieve p he will try again, but if he does not achieve q, he will not necessarily try again. r is the argument to which the P-R-GOAL is relative, the other mental states conditionally to which the agent adopts the goal.

C&L Def 15:

$$\{\text{REQUEST } x\ y\ e\ a\} = \text{def } \{CA\ x\ e\ [(\text{BMB } y\ x\ (\text{GOAL } x\ \Diamond(\text{DONE } y\ a))) \wedge (\text{P-R-GOAL } y\ (\text{DONE } y\ a)\ (\text{GOAL } x\ \Diamond(\text{DONE } y\ a)))] [(\text{HELPFUL } y\ x) \wedge q \wedge \Diamond(\text{DONE } y\ a)]\ r\}$$

A request is now defined as a committed attempt by the speaker x, to bring about the state of affairs where it is mutually believed between the speaker and hearer that the speaker wants the hearer to perform the requested action. He also intends the hearer to adopt a persistent goal to perform the action as long as it doesn't conflict with one of his own, relative to the fact that the speaker wants him to. The speaker's chosen effects for all this are that he wants the hearer to be helpful and therefore take on the

P-R-GOAL and not drop it for any other reason than that it is achieved. The hearer should therefore eventually do the act.

Cohen and Levesque (1987b) point out that the speaker need only believe and intend the act to make the conditions true; he does not need to actually make them true. Also, since there is no indication how the hearer arrives at the mutual belief that the speaker wants the hearer to do some action, this definition is equally appropriate for indirect as well as direct speech acts.

The definition of request given above refers to the event e. It allows for performing any sequence of actions that produce the required effect, for example, the uttering of an imperative. This is done by uttering a sentence which must fulfill certain preconditions, these being that x be the agent of e , y be attending to x, e be the event of x uttering sentence s to y, the sentence s comprise an imperative, and it is not the case that at any level of alternating belief between x and y that the speaker is not sincere about having a goal for y to do something.

In the definition of a request given above, and also in those offered by Perrault (1987) mutual beliefs are an intended effect of the speaker. This is another difference between Cohen and Levesque's (1987b) and Perrault's (1987) analyses, and that of my framework. From C&L Def's 7 and 8:

$$(\text{BMB } H\ S\ (\text{GOAL } S\ \Diamond p)) \supset (\text{BEL } H\ (\text{GOAL } S\ \Diamond p)) \wedge (\text{BEL } H\ (\text{BEL } S\ (\text{GOAL } S\ \Diamond p)) \wedge (\text{BEL } H\ (\text{BEL } S\ (\text{BEL } H\ (\text{GOAL } S\ \Diamond p)))) \text{.....and so on.}$$

In other words, if it is mutually believed between H and S that S has a goal that eventually p, then H believes S to have this goal, and believes that S believes she has this goal too. In this thesis' strategic framework, cooperative and autonomous agents understand their communicative action as conforming to a strategic rationality. Cohen and Levesque's mutual beliefs are replaced with a belief about the speaker's interests and expectations, or good strategy, on the part of the hearer. This means that in addition to believing the speaker has a communicative goal, the hearer believes the speaker believed this goal would

be successful, and the desired belief or goal therefore adopted by the hearer. This was expected by the speaker to then lead to subsequent action on the part of the hearer, in her favour. A goal recognised whilst believing that the speaker would have considered that the hearer would not adopt it, might express a joke, or irony, for example. Elaboration of these ideas, and a more detailed contrast with Cohen and Levesque's approach is found in the latter part of the next section.

6.3.1.2 The strategic approach

Acquiring a mental state as a result of another's dialogue action is treated here as a consequence of commitment to a goal by the hearer, in the same way as the performing of some act is a consequence of a commitment to a goal. Agents have to generate P-R-GOALs in order to eventually adopt a belief. For example,(P-R-GOAL x (BEL x q) p) will lead to ◊(BEL x q) if this is not already achieved, impossible to achieve and as long as the conditions p, are true. Previous work in agent theory neither contradicts such an approach nor supports it. The psychological validity has not been considered, given that the aims of this research are to develop a model for future machine applications.

Belief and goal acquisition during dialogue requires the hearer to generate a goal to adopt the speaker's goal that the mental state in question be acquired. Incorporating some of the notions discussed in chapter 2, the suggested conditions for adopting another's goal that a mental state be acquired by oneself, as a result of receiving or apprehending a dialogue action, are as follows:

firstly, recognition of another agent's goal to induce a particular mental state. This will result in the agent's adoption of that goal if it has not already been adopted and secondly:

if the goal concerns the adoption of a belief:

there is evidence that the speaker knows the proposition he wants the hearer to adopt as a belief, and therefore it must be true.

OR

if the goal concerns the adoption of a belief or a goal:

adopting it is preferred to not adopting it.

An example of a belief being adopted following recognition of another's communicative goal to do so, and on the basis of evidence in the world would be: A has a goal for B to believe it is raining outside, and B recognises A's goal as well as noticing he has entered the room with a wet umbrella. B believes that A knows it is raining. However, if B believes A knows it is raining, she does not need to also be told this. She will inevitably adopt the belief state, regardless of the speech action:

Theorem 1

$(BEL\ x\ (KNOW\ y\ p)) \supset \Diamond(BEL\ x\ p)$

This says that if agent x believes another agent y to know p, there being evidence of its truth in the world, then x eventually believes p.

Proof:

1. $(BEL\ x\ (KNOW\ y\ p))$

x believes y knows p.

2. $(BEL\ x\ ((BEL\ y\ p) \wedge p))$ C&L: Def 5

According to the definition of KNOW, if x believes y knows p then x believes that y believes p, and p is true.

3. $(BEL\ x\ (BEL\ y\ p)) \wedge (BEL\ x\ p)$

4. $(BEL\ x\ p)$ C&L Proposition 5

Believing p extends into the future from "now" (Cohen & Levesque,1987a).

5. $\Diamond(BEL\ x\ p)$

What we want is a statement about the aquisition of a belief state as a result of a dialogue action and in the absence of evidence of the proposition's validity in the real world. In such cases the receiving agent is faced with a set of circumstances for which she has a preference. Assumption 3 (pp139-140) states that following recognition of the communicative goal, she believes she will eventually have adopted the belief or goal, or she will not. From Assumption 1 (pp110), she can generate a goal for one of these options to be the case. This is described in theorem 2:

Theorem 2:

$(BEL\ x\ (GOAL\ y\ \Diamond q)) \wedge (PREFER\ x\ q\ \sim q) \supset (GOAL\ x\ \Diamond q)$

This says that if agent x recognises y's goal q, and x prefers q than not q, then x adopts the goal q.

Proof:

1. $(BEL\ x\ (GOAL\ y\ \Diamond q))$ Given

x recognises y's communicative goal q.

2. $(BEL\ x\ (\Diamond q \vee \Diamond \sim q)) \wedge (PREFER\ x\ q\ \sim q)$ Assumption 3

x believes either eventually q or eventually not q will be the case, and also prefers q to not q. Preferring is defined in Def 1 (pp110) as the agent having a belief that if she could eventually believe one term of a disjunction between two options then there is one which she would have a goal to eventually believe rather than the other. In this case this is q rather than not q.

3. $(GOAL\ x\ \Diamond q)$ Assumption 1.

From Assumption 1, if the agent has a preference between two options, and is faced with these , a goal for the preferred option results.

An example : y may say "Please open the door" with the goal that x generate a goal to open the

door, presumably recognising y's plan to pass through it. x will adopt that goal as long as she has a belief that if faced with the alternatives of opening the door for y in these circumstances, or not opening the door for y, she will generate a goal to open the door for y. Thus, she is not merely being benevolent, but adopts the goal on the basis of being faced with a choice for which she personally has a preference:

$$(BEL\ x\ (GOAL\ y\ \Diamond(DONE\ x\ a))) \wedge (PREFER\ x\ (DONE\ x\ a)\ \sim(DONE\ x\ a)) \supset (GOAL\ x\ \Diamond(DONE\ x\ a))$$

a represents opening the door. Agent x recognises y's goal for x to have done a, and x prefers the act having been done to not having been done. x adopts the goal to have done a.

It is most likely, that x has other goals to which she is committed at the time. Perhaps she is going to get something, she is walking , and so on. If the consequences of adopting y's goal are that she cannot continue getting what she was on her way to get for instance, then there is an existing contradictory goal. By preferring to open the door for y than not, she is preferring this to that original goal. If she then opens the door, she must have dropped the original goal. This is only possible if she really does have two alternatives from which to choose :

$(BEL\ x\ (\Diamond(DONE\ x\ a) \vee \Diamond\sim(DONE\ x\ a)))$ from assumption 3. There must therefore be an understanding that existing goals or beliefs can be altered; their contradiction cannot be believed to be forever false. If x currently believes p or has a goal to have done a, then there are conditions under which x can drop that belief or goal. These are described in the following:

Assumption 4:

$$(BEL\ x\ (BEL\ x\ \sim p)) \wedge (BEL\ x\ (GOAL\ y\ \Diamond(BEL\ x\ p))) \supset (BEFORE\ (PREFER\ x\ (BEL\ x\ p)\ (BEL\ x\ \sim p))\ (BEL\ x\ \Diamond(BEL\ x\ p)))$$

If x believes she believes not p and recognises another's goal that eventually she believe p, then this will only be the case if she prefers to believe p to not p.

or

$$(BEL\ x\ \neg\Diamond(DONE\ x\ a))\ \wedge (BEL\ x\ (GOAL\ y\ \Diamond(DONE\ x\ a))) \supset$$

$$(BEFORE\ (PREFER\ x\ \Diamond(DONE\ x\ a)\ \neg\Diamond(DONE\ x\ a))\quad (BEL\ x\ \Diamond(DONE\ x\ a)))$$

x believes she will not eventually have done a, but recognises another's goal that she does. The only condition for her changing her belief is if she now prefers a to not a , the preference being ascertained in the existing circumstances.

In both these cases, having the belief also implies having the goal from C&L: proposition 3:

$(GOAL\ x\ \Diamond(BEL\ x\ p))$ and $(GOAL\ x\ \Diamond(DONE\ x\ a))$

Another example. Othello initially believes Desdemona to be a faithful wife. He also believes Iago to be a loyal and trustworthy friend. In conversation with Iago, Othello recognises Iago's goal of inducing a belief in him that Desdemona has not been faithful. Iago reasons that Othello will adopt this belief and reject the existing contradictory one, partly because his beliefs about Iago's honesty and friendship are stronger/ more central than his beliefs about Desdemona's love; he prefers to believe Iago is honest than Desdemona is faithful. Once having recognised Iago's communicative goal, he can no longer believe Iago to be honest and Desdemona faithful. His preference is therefore to adopt the belief that Desdemona has been unfaithful, but this is only in the light of Iago's action. Iago also ensures the adoption of this belief by Othello at a later stage, with the introduction of evidence of its truth in the world in the form of a handkerchief.

In the example above, Othello is faced with choices between Iago's honesty and Desdemona's fidelity only in the light of Iago's communication. In dialogue, new information is being presented which alters the context for determination of preference from those previously determined. A goal to go swimming for example, may have been made originally because of a preference to be swimming than not, and all that "not" implied in terms of other goals satisfied. However, in the light of new information about a possible alternative, not swimming has different implications or consequences. The preference being ascertained is a different one.

The conditions described so far in this section for adopting another's goal that a particular mental state be acquired following the recognition of this as a communicative goal, do not question the validity of the recognised goal. A goal of the speaker's to induce a belief in the listener having been recognised, it is considered in relation to the conditions of either evidence of truth of that belief in the world, or having a preference. In the latter case, the listener then either chooses to adopt or not to adopt the belief. If the belief is to be adopted, the listener generates a goal to this effect, and has therefore adopted the speaker's goal as her own. It may be the case, however, that the hearer recognises a goal for the speaker to believe p when this was not the desired effect at all; it was some other proposition q that the speaker had a goal for the hearer to believe. Perhaps a speaker has a goal for the hearer to recognise his goal for her to believe p, yet he does not really expect her to adopt it, as for example in a joke.

In strategic interaction the communicative actions of autonomous agents are understood as conforming to a strategic rationality.

The notion of dialogue as strategic interaction and dialogue actions therefore conforming to a strategic rationality, was introduced in section 3.2.5 , and section 3.2.5.2 in particular. Autonomous, rational agents were described as understanding themselves and others to perform dialogue actions on the basis of firstly, the mental state they want to induce in the hearer, secondly, whether they believe this mental state will be achieved, and thirdly, their subjective expectations of subsequent dialogue action on the part of the hearer being also in their interests. These ideas are formally expressed in the definition of the predicate Gd-STRAT, defined below. The hearer's beliefs concerning the speaker's considerations of

good strategy are important to the hearer's decision whether to adopt another's intended mental state in dialogue. It assists the interpretation of the utterance, and the detection of non-serious, or ironical utterances. An example was given at the end of section 3.2.4, whereby a hearer recognising the speaker's goal that she jump in a lake, doesn't believe that the speaker believed she would actually jump in a lake. Since the action the speaker performed must have conformed to a strategic rationality, there must have been some other goal she was expected to adopt. This then, is another important condition for acquisition of a desired mental state recognised as another's goal in dialogue. Acquiring a mental state is a result of commitment to the goal to do so, and this commitment is relative to the hearer believing that the speaker performed her communicative action considering it to be a good strategy for her.

Assumption 6:

$$(\text{BEL } x\ (\text{GOAL } y\ \Diamond q)) \wedge (\text{PREFER } x\ q\ \sim q)$$

$$\supset\ (\text{P-R-GOAL } x\ q\ (\text{BEL } x\ (\text{Gd-STRAT } y\ x\ \Diamond q)))$$

q represents a mental state of x's; for example (BEL x p) or (DONE x a). Assumption 6 says that if agent x recognises y's goal as q and prefers q to not q, then x will generate a persistent goal to achieve q, relative to believing y considered q a good strategy for y.

A persistent relativised goal can only be dropped for one of three reasons. These are: because it is achieved, believed impossible to achieve, or the reason to which it is relative is not true. A persistent goal dropped for no reason other than that it is achieved, is eventually achieved (from C&L Theorem 2 as described in section 4.4.3.2). **Assumption 6 therefore describes the conditions whereby a mental state can be acquired during dialogue in the absence of any evidence of its truth in the world.** It should be noted that the hearer's commitment to the goal to acquire a mental state during dialogue is relative to a belief that the speaker considered it a good strategy for him. This means that misunderstandings can still occur, because q may not have been what y had a goal for x to adopt at all, even though x eventually adopts it believing it to be what y considered his good strategy. Good strategy is defined as follows:

Def. 13: (Gd-STRAT x y p) = *def* (INTERESTS x p)

$\wedge$ $\forall q \forall s \forall e$ (p?;e;q?) (INTERESTS x (CA y e q s r))[1]

Good strategy for x with y in relation to p is defined as x believing that p is in her interests, and for all states q and s resulting from events e, that would follow the successful attainment of p, it is also in x's interests that y makes a committed attempt by event e to achieve q. This latter half of the conjunction simply means that the expected subsequent action which x considers y may intend to perform, should also be in x's interests.

Gd-STRAT is recursive with no stopping condition. This is because in the above definition, in $\forall q \forall s \forall e$ (p?;e;q?) (INTERESTS x (CA y e q s r)), r is (Gd-STRAT y x q). In other words, determining the subsequent action of the other agent y, involves assessments of their assessments of whether x believes y will consider q to be a good strategy. It is an infinite regression. What is taking place looks like this:

INTERESTS (α) AND EXPECTATIONS (β)

⇓

INTERESTS (β) AND EXPECTATIONS (α)

⇓

INTERESTS (α) AND EXPECTATIONS(β)

⇓

.....etc.

[1]CA is defined as C&L Def 14 in section 6.3.1.1.on pp 144

In other words, a committed attempt by α to achieve some effect, is relative to:

(i) the effect being in α's interests. Reasoning about this involves looking at α's own goals and beliefs concerning whether the other agent involved β, will achieve this effect, and

(ii) the expected subsequent state which β will want to induce as a result of α 's action being successful, being also in α's interests. Now, in order to assess this, the same basis must be used because β will be generating his action according to his own interests and expectations of α adopting the desired state and subsequent action, and so on.....

The suggestion for managing this infinite regression is to place an artificial limit on the number of recursions that can take place . The asssumption is that reasoning about appropriate dialogue action includes : "well, if I say a, then she might say b, and I want her to say b " , but that practical reasoning about action does not really require " well, if I say a, then she might say b, believing that I would then say c, ...etc. " It is believed that the computational effort in such calculations would outweigh the possible benefits. Inserting stopping conditions into situations where recursions should theoretically continue into infinity has been the practical solution adopted by other researchers, some in different fields of study. This has previously been discussed in section 3.2.1 with respect to definitions of mutual belief and mutual knowledge. Also Howard's work on metagames (Howard, 1971) discussed in section 3.2.5 limits levels of metagames to two, when determining the most rational move in a two-person game.

The role of good strategy in assumption 6 can be illustrated with an example. x recognises y has a goal for her to have done some act a: (BEL x (GOAL y ◊(DONE x a))). x takes on the goal to have done a with commitment as long as she prefers to have done a than not a, and also believes that y does have a goal for her to do a which he also expected she would adopt. She also believes that y believed that any subsequent action of hers would be favourable to him. If x does not believe y would have considered x to have adopted ◊(DONE x a) as her own goal for example, then she does not believe ◊(DONE x a) was a good strategy for y. Perhaps a represents something like walking off a cliff or jumping in a lake, as in the example referred to earlier. Then ◊(DONE x a) must not have in fact been

the goal for recognition. But, dialogue actions are understood as having been generated according to a strategic rationality. There must be some other goal she was expected to recognise and adopt. Once established, this will be adopted by her as a committed goal, only to be dropped once achieved or impossible to achieve, just as long as it is also preferred to its contradiction.

In this example, the issue seems to be whether the hearer believes the speaker is "sincere", using Cohen and Levesque's (1987b) terminology. Adopting the recognised goal is seen by them as dependent upon more simply whether the hearer believes the speaker also believes she has this as a goal, and that she does have this as a goal. To illustrate the point with another example: if A requests B to make a cup of tea, B recognises A's goal for B to believe A wants B to make them the tea, but B will only interpret it as a request and eventually make the tea, if she believes A really does have a goal for tea. Cohen and and Levesque use "Go jump in the lake" as their example (Cohen and Levesque, 1985,1987b). They suggest that the request is recognised, but if the hearer believes the speaker does not really want the hearer to freeze by jumping in a lake or there is no lake in the immediate vicinity, then the request interpretation is blocked.

The strategic approach however, reasons about dialogue action on the basis of the recognised communicative goal, the hearer's assessments of the speaker's expectations concerning whether the intended effects on the hearer's mental states will be successfully effected, and the hearer's subsequent action if it is. Are there any other gains for this extra complexity? The following scenario and examples will be used to demonstrate:

The context is one of a colleague or relative having popped in for a brief visit. After announcing that she must leave, the hostess says : "Oh! Do you have to leave so soon?" According to Cohen and Levesque's approach, the visitor v, will recognise the hostesses h, goal for her to believe she wishes her to stay a bit longer. Assuming sincerity, v believes h has intended a mutual belief that there is a goal for her to stay. Being helpful and having no contradictory goal, v will stay. However, such a request could be made merely out of politeness and with no real desire for the visitor to stay any longer than necessary. This example is in no way as clear cut as there being no lake around in the "Go jump in the lake example". With Cohen and Levesque's framework, v would need to recognise h's insincerity to recognise it as a non-serious request. But on what basis can she infer h's goal as one other than what

was stated? In order to satisfy the goal of politeness, the hostess cannot make it too obvious that she wants to be left in peace.

With a strategic approach there is a basis upon which reasoning about h's goal can occur. v knows that the action was generated relative to her interests and expectations of subsequent action. Therefore, firstly h must have a goal for v to believe she has a goal for her to stay. Did she also however, believe that v would adopt the goal to stay, as her own? If v believes that h expected her not to take on the goal, perhaps on the basis of past experience of such situations where v has never stayed, or v knowing that h knows that v has another engagement, then v can reason that this is not really a request to stay. h must have also considered that v's subsequent action should hopefully be in h's interests. However, was it v's staying which was really the goal or was it v believing h is polite and hospitable? Of course, v might still misunderstand in a situation like this one. However, at least there is some basis upon which an assessment can be attempted. Cohen and Levesque's framework would only allow the hearer to acknowledge "non-serious requests" (Cohen and Levesque, 1987b) in situations where there was little ambiguity.

Reasoning about irony can similarly be potentially successfully achieved. Perrault offers "This was a wonderful meal" as an example of an utterance intended to ironically convey the belief that the speaker did not consider the meal to be wonderful (Perrault 1987). His analysis of irony is as follows: the speaker asserts p but it is mutually believed that before the utterance was made, neither of them believed p. Given that H assumes that S cannot adopt a new belief contradictory to an existing and therefore persisting one because of Perrault's persistence theory of belief (as described in chapter1), S must be being ironical. This analysis does not explain why he couldn't have been lying, or even have changed his mind (Morgan 1987). On the other hand, with the assumption that S performed the utterance believing it to be a good strategy for him, it is the belief that the hearer was not expected to adopt a belief that the speaker believed the meal to be good, that blocks the interpretation as a straightforward and sincere assertion. This belief would presumably be acquired also as a result of the communication, via appropriate facial expressions or intonation perhaps (Morgan 1987).

A lie can be successfully carried out by the hearer recognising the speaker's communicative goal and believing the speaker to have believed he would adopt the intended mental state, and subsequently act in his favour. An unsuccessful lie occurs when the hearer chooses not to adopt the speaker's intended

mental state on the basis of the conditions discussed earlier, of evidence or preference.

As a final comment in this section concerning the strategic basis to belief and goal acquisition in dialogue, it is interesting to note that assumption 6 is very similar to Def 10 in section 5.4. This is the definition of COOP-1 as the type of cooperation whereby one agent has a goal and the other adopts it on the basis of her own preferences, and also because the other has it as a goal. Adopting another's communicated goal in dialogue to acquire a mental state has the same conditions; firstly preference, and secondly the other having this as a goal is a component of believing the other to have considered it as a good strategy. It seems intuitively correct that between autonomous agents, the acquisition of a mental state in dialogue as a result of another agent's goal that one do so, be a cooperative act.

6.4 Control over information revealed in dialogue

Autonomous agents have control over the information they acquire in dialogue. They have control over the information they receive; they have control over their own mental states. This control over one's own mental states extends also to information revealed in dialogue, assuming dialogue as intentional communication. Unintentional communication on the other hand, conveys information regarding one's mental states but with no control. For example:

> "A can convey that he is afraid, nervous or upset by shaking hands with a sweaty and trembling hand" (Allwood 1976).

Dialogue as intentional communication is a conscious manipulation of others' mental states. According to Grice (Grice 1957, 1969), in order for the hearer to correctly interpret the meaning of an utterance, she must recognise the desire to perform these manipulations. The utterance must convey to the hearer that the speaker has an intention (goal) to induce a particular mental state in the hearer. For example, if x has a goal that y believe that p, then x may perform some utterance. Whether it is a straightforward assertion that p, or a question comprising some indirect speech act, it will have properties which enable

y to recognise the goal. Perhaps the goal is arrived at as a component of a recognised plan. A belief such as: (BEL y (GOAL x ◊(BEL y p))) results. The speaker also has another goal that the hearer should recognise her goal: (GOAL x ◊(BEL y (GOAL x ◊(BEL y p)))). Conveying such information relies not only on certain features of the utterance itself, but also on the hearer's beliefs and the context in which the interaction takes place.

Agents can sometimes unintentionally convey information at the same time as doing so intentionally. Saying "How nice to see you" with a tired and hollow smile on one's face, for example. The speaker's goal may be to effect a belief in the other agent that she is pleased to see her, and the hearer will recognise this goal and form the corresponding belief : (BEL x (GOAL y (BEL x y-pleased-to-see-x))). In other words, she believes that y wants her to believe this. However, in this case the tired and hollow smile conveyed unintentionally, leads to another belief: (BEL x (BEL y ~y-pleased-to-see-x)). What is important to note in this example, is that whether x forms the second belief or not, she has understood the communication; she has formed the appropriate belief regarding y's intentional state, which is the first belief. The additional information as to the speakers <u>own</u> relation to the propositional content of the desired mental state, was an extra, available here by unintentional communication, but was unnecessary to perceiving y's communicative intentions (goal) and therefore in the correct interpretation of the communication. This view is confirmed by Bach (1987) and Bach & Harnish (1979).

> ".....to make an utterance with a communicative intention is to express an attitude. He may or may not actually possess it, but that question is irrelevant to the purely communicative aspect of his utterance." (Bach 1987).

6.4.1 Expression - a relation between an utterance and a mental state

In intentional communication, an utterance expresses the fact that the speaker has a goal that a particular belief or goal be induced in the hearer. The only state required to be recognised for effective

communication is that speaker's goal regarding the receiver's mental states. The belief or goal to be induced in the hearer has various relations to the speakers own beliefs and goals, and what these relations are may be inferred for future action, but are not actually necessary for the communicative process. The speaker has control over her mental states. The speaker may herself hold the belief or goal she wants to induce in the hearer, in which case the speaker believes p for example, and has a goal for the hearer to believe p. If the opposite were the case, the speaker whilst believing p, would have a goal for the hearer to believe not p. In either of these cases, the utterance can also be concealing. This means that the speaker has a belief with respect to p but has a goal for the hearer to not have a belief about the speaker's belief with respect to p. Alternatively, it could be revealing, and the speaker therefore have a goal for the hearer to have a belief about the speaker's belief with respect to p.

This relation between an utterance and the speakers beliefs is that the utterance is an expression of those beliefs. The nature of the relation is of different types, these being veracious expression, mendacious expression, revealing expression or concealing expression of the agents beliefs. For example, (VERACIOUS e (BEL x p)) says that the utterance e is veracious with respect to x's belief that p. (MENDACIOUS f (BEL x p)) says that the utterance f is mendacious with respect to x's belief that p. The different expressions or types of relation between an utterance and the speakers beliefs, are defined in the following section.

6.4.1.1 Definitions of veracity, mendacity, revealing and concealing

Veracity, mendacity, concealing and revealing are defined as action expressions whereby the utterance or event is done following a belief that a mental state will thereby be recognised by the hearer. What distinguishes these expressions is the relation between the speaker's true mental state and that which she believes will result in the hearer after the event is done.

Def.14: (VERACIOUS e (BEL x p)) $=_{def}$

$$[(\text{BEL } x\ (\text{HAPPENS } x\ e;\ \forall y(\text{BEL } y(\text{GOAL } x\ \Diamond(\text{BEL } y\ p)))?)) \wedge (\text{BEL } x\ p)]\ ?;\ e$$

A veracious event with respect to x's belief p, is one which is done when believing that for all agents y,

doing e will achieve y recognising x's goal that y believe p, and x believes p.

Likewise, (VERACIOUS e (GOAL x ◊(DONE y a))) =*def*

[(BEL x (HAPPENS x e; ∀y(BEL y (GOAL x ◊(DONE y a)))?))

∧ (GOAL x ◊(DONE y a))] ?; e

A veracious event with respect to x's goal that y have eventually done a is one which is done when believing that for all agents y, doing e will achieve y recognising x's goal that y have eventually done a , and x does have the goal that y have eventually done a.

Def.15: (MENDACIOUS e (BEL x p)) =*def*

[(BEL x (HAPPENS x e; ∀y(BEL y(GOAL x ◊(BEL y ~p)))?)) ∧ (BEL x p)] ?; e

A mendacious event with respect to x's belief p, is one which is done when believing that for all agents y, doing e will achieve y recognising x's goal that y believe not p, and x believes p.

Def. 16: (CONCEALING e (BEL x p)) =def

[(BEL x (HAPPENS x e; ∀y ~◊(BEL y (BEL x p))?)) ∧ (BEL x p)] ?; e

A concealing event with respect to x's belief p, is one which is done when believing that for all agents y, doing e will result in y not forming a belief with respect to the speaker's belief p. This may operate in conjunction with a veracious or mendacious act related to the same or a different belief. Alternatively, e may be silence.

"Lying is done with words, and also with silence" (Rich, 1975).

Def. 17: (REVEALING e (BEL x p)) =*def*

[(BEL x (HAPPENS x e; ∀y(BEL y (GOAL x ◊(BEL y (BEL x p))))?))] ?; e

A revealing event with respect to x's belief p, is one which is done when believing that for all agents y, doing e will result in y recognising the speaker's goal that y form a belief about the speaker's belief p.

A mendacious utterance may also be a concealing utterance, in which case it is a lie aimed at being kept undiscovered. A mendacious utterance which is revealing, expresses irony. Veracious utterances can also be either concealing or revealing. An utterance aimed at inducing a belief p, in another, may or may not also convey the information that the speaker wants the hearer to form a belief that the speaker also believes p.

6.4.2 Practical implications

x's own beliefs about p have been stated as irrelevant to y's understanding of an utterance which results in: (BEL y (GOAL x ◊(BEL y p))), for example. However, actually determining by inference a belief about the speaker's beliefs with respect to p, may be important. Such inferences may affect the hearer's choice to adopt the speaker's desired mental state. They may be necessary for reasoning about subsequent generation of appropriate dialogue action. Explicit representation of the various possible expressions of utterances are therefore practically important.

Assumption 7:

(BEL x (GOAL y ◊(BEL x p))) ⊃ (BEL x ((BEL y p) ∨ (BEL y ~p)))

This says that on recognition of another's goal that the hearer eventually believe p, the hearer believes that the speaker has some belief about p, either p or not p.

Inference by the hearer as to which of these is believed to be the case requires also other relevant beliefs, such as those derived from previous experience of the speaker, and the posture of the agents with respect to the proposition p. In addition, a default assumption is necessary, that truthfulness is assumed in the absence of any inferred or existing beliefs which may suggest the contrary. The issue of assumed tendencies towards truth-telling in the light of potential further interactions, was discussed in sections 3.2.4. and 5.7.2. It relates to assumed mutual beliefs between interacting agents, regarding various detrimental effects of deception on future interactions. The principle was stated in chapter 3 as follows:

agents believe themselves and others to be able to choose to perform dialogue actions which are mendacious and/or concealing expressions of a mental state, but agents are assumed to use revealing and veracious expressions unless the receiving agent has any beliefs which might suggest the contrary.

Assumption 8:

$$(BEL\ x\ (GOAL\ y\ \Diamond(BEL\ x\ p))) \wedge \sim(\exists q\ ((BEL\ x\ q) \supset \sim(BEL\ x\ (BEL\ y\ p))))$$

$$\supset\ (BEL\ x\ (BEL\ y\ p))$$

This says that if x recognises y's goal that she believe p and x has no other belief which would imply that x should not believe that y also believes p, then x believes y believes p.

6.5. Communicative acts

A communicative act is characterised in the following sub-section, to demonstrate the integration of the principles of autonomy and strategic interaction detailed in the previous sections 6.3. and 6.4. The act chosen is a request, as a means of directly contrasting in sub-section 6.5.2, with Cohen and Levesque's characterisation of a request.

6.5.1 The strategic approach

The strategic generation of a communicative action such as a request, does not require assumptions that the hearer will helpfully adopt the speaker's goal just because he has it. As stated in Assumption 6 which was described in section 6.3, adoption of the goal is conditional upon the hearer's recognition of the speaker's goal, the hearer's own beliefs and preferences, and assessments of the speaker's considerations whether he would adopt the goal and subsequently act in the speaker's favour. This means the hearer is not expected to be benevolent, nor believe the speaker to be necessarily sincere.

Def 18: {REQUEST x y e a} = def

$$\{CA\ x\ e\ [(BEL\ y\ (GOAL\ x\ \Diamond(DONE\ y\ a))) \wedge$$

$$(P\text{-}R\text{-}GOAL\ y\ (DONE\ y\ a)\ s]$$

$$[t \wedge v \wedge \Diamond(DONE\ y\ a)]\ (Gd\text{-}STRAT\ x\ y\ \Diamond(DONE\ y\ a))\ \}$$

The request is made as a committed attempt (C & L Def 14, p144) to achieve the state of affairs whereby the hearer y, recognises the speaker's goal for her to do some act and she adopts this goal as a persistent one. The persistent goal is relative to believing that the speaker x, considered it a good strategy for him. The chosen effects are that t will be true and v and therefore y will eventually have done a. The persistent goal for all this is relative to x's consideration of ◊(DONE y a) as a good strategy.

t = def $(BEL\ y\ (GOAL\ x\ \Diamond(DONE\ y\ a))) \wedge$

$(PREFER\ y\ (DONE\ y\ a)\text{~}(DONE\ y\ a))$

y recognises x's goal for y to have done some act a. Also she prefers to have done a than not done a. These are the conditions in Theorem 2 which if true, then y adopts x's goal: (GOAL y ◊(DONE y a)).

s = *def* $(BEL\ y\ (Gd\text{-}STRAT\ x\ y\ \Diamond(DONE\ y\ a)))$

y believes that the speaker x, considered it a good strategy for him that y should eventually have done a.

v = def $\text{~}[BEFORE\ [\text{~}(COMPETENT\ y\ (DONE\ y\ a)) \vee (BEL\ y\ []\text{~}(DONE\ y\ a)) \vee \text{~}s\ ]$

$\text{~}(GOAL\ y\ (LATER\ (DONE\ y\ a)))]$

y's goal to have done a is only dropped because it is achieved. It was not dropped because y was not competent to achieve it, y believed it to be impossible, or y did not believe that x considered it a good strategy for x.

As with C&L's definition, there is no indication of how the hearer will form the belief:

(BEL y (GOAL x ◊(DONE x a))); whether directly or indirectly.

A request has been strategically defined above as a committed attempt to achieve a particular state of affairs which relates to the mental state of the hearer. Other communicative acts can be defined similarly as complex action expressions to achieve different effects on the mental state of the hearer. Incorporating the ideas comprising section 6.4, which relate to agents' control over the mental states they reveal, utterances can be defined also as veracious or mendacious, and concealing or revealing expressions of the speaker's true mental states. For example, from Def 18. and Def 14, a sincere request is defined as

Def 19:

$\{$V-REQUEST x y e a$\}$ = def

$\{$CA x (VERACIOUS e (GOAL x $\Diamond$(DONE y a)))

$$[(\text{BEL y (GOAL x } \Diamond\text{(DONE y a)))} \wedge \text{(P-R-GOAL y (DONE y a) s)}]$$

$$[t \wedge v \wedge \Diamond\text{(DONE y a)}]\ \text{(Gd-STRAT x y } \Diamond\text{(DONE y a))}\ \}$$

Here, the request is a complex action expression to achieve the desired state of affairs, and comprises an event which is done when believing y will recognise her goal for y to do a and when x does in fact, have the goal $\Diamond$(DONE y a).

Def 14:

(VERACIOUS e (GOAL x p) = def

[(BEL x (HAPPENS x e; $\forall$y (BEL y (GOAL x p)))?)) $\wedge$ (GOAL x p))]?; e

6.5.2 The strategic approach contrasted with Cohen and Levesque's - an illustrative summary

In section 6.3.1.1, the Cohen and Levesque definition of a request is described as a committed attempt on the part of the speaker S, to achieve the following :

(i) a mutual belief that S has a goal. For example, a goal that H eventually have done a,

(ii) H to adopt and be committed to this goal because S has that goal and H has no existing contradictory one; H is helpful,

(iii) H to only drop this goal because it has been achieved and therefore,

(iv) H to have eventually done a (Cohen & Levesque, 1987a).

In contrast, the strategic definition of a request is described in Def. 18 as a committed attempt on the part of the speaker S, to achieve the following:

(i) a belief in H that S has a goal. For example, a goal that H eventually have done a,

(All that is required here is that H recognise S's communicative goal, not that it be a mutual belief.)

(ii) H to adopt this goal in the light of S's goal being recognised, and having done a being preferred by H than not having done a

(Not benevolent, but autonomous goal adoption.)

(iii) H to be committed to this goal because he also believes that S requested it on the basis of S considering it a good strategy for him

(No assumed sincerity. Dialogue as strategic interaction assumed.)

(iv) H to only drop this goal because it has been achieved and therefore,

(v) H to have eventually done a,

(vi) and all this is relative to S believing H eventually having done a and any subsequent action of H's, to be in S's interests; it is a good strategy for S.

(Dialogue as strategic interaction assumed.)

6.6 Strategic objectives

In section 6.5, an example communicative or dialogue act was defined appropriately to the notion of dialogue as strategic interaction between autonomous agents. According to the views expressed throughout this thesis, such an act is the means of potentially manipulating another's mental states and consequently achieving a desired posture. The discussions so far have all started with the premise that there is a particular goal state which relates to the hearer's mental states, which the speaker has a goal for the hearer to recognise. It was suggested early on in chapter 1 that this goal might be a component of a plan - a sub-goal to another goal of the interaction. Apart from this, nothing has been said in this thesis about determining goals; agents just have them. Nothing has been said about determining which of alternative potential goals to become committed to achieving, or in what order if there are more than one and they cannot be achieved concurrently. Resulting from the proposed theory of cooperative dialogue as strategic interaction, there are however, two overall strategic objectives according to which particular goals may be selected as those which the agent will adopt with commitment. These are described as a result of a brief review of the nature of strategic interaction, as follows:

In strategic interaction, agents assume themselves and each other to perform dialogue actions as the result of having a persistent goal which is relative to a desired effect on the other's mental states being a good strategy for the speaker. Assessments of good strategy are therefore fundamental components of strategic interaction, as expressed by the performance of dialogue actions. In the definition of good strategy, both the effect under consideration and any subsequent response on the part of the hearer are described in terms of these being in the speaker's interests. As fundamental components of good strategy, interests are therefore most central to reasoning about dialogue actions.

Interests were described in chapters 2 and 4 as being comprised of two parts; firstly the goal or desired effect on the hearer's mental states and secondly, a belief that such a goal will be achieved. This means that dialogue action generation is reasoned relative to these two considerations. In addition it means that another's subsequent action is also reasoned about in relation to both of these. Strategies for dialogue action generation as a means of potentially affecting another's assessments of their own good strategy and correspondingly their future action, can therefore be determined according to two alternative

objectives. The first relates to the other's believed goal. The strategic objective may be to encourage or discourage the other's believed goal. For example, x has a goal for y and x not to go to see the film "Angel Heart". x believes y has a goal for y and x to go to see the film "Angel Heart". Both have persistent goals to change each other's goals; there is a conflict between x and y with respect to them going to see this film. A possible good strategy for x is to attempt to induce a belief in y that the film contains horrific and disturbing sequences, based upon x's beliefs concerning y's preferences related to films involving gory effects, for example. The consequence of this belief if successfully induced, is predicted by x as being that y would no longer have a goal to change x's goal.

The second strategic objective relates to the likelihood of the goal state being achieved. For example, inducing a mental state in the other agent such that she will not believe that her desired mental state in oneself will be achieved. In the example above, such a good strategy might be for x to attempt to induce beliefs in y that "Angel Heart" is only currently showing in a cinema out of town, and that x has absolutely no desire to travel out of town. Subsequent to the successful attainment of this goal state, y may still have the goal for x to also have a goal for them to see this particular film, but he believes that x will not actually take the goal on.

The alternative strategic objectives therefore concentrate on either encouraging or discouraging the two components of interests relative to which an agent's commitment to future action is determined. These are firstly, the other agent's believed goal, and secondly, the believed likelihood of successful goal attainment. The basis for prediction in both is an understanding of the conditions under which mental states are adopted in dialogue. These were described earlier in this chapter as evidence or preference, and predictions about the speaker's assessments of this as a good strategy for them.

6.7 Conclusions

In this chapter, agents have been defined as autonomous in dialogue, on two counts. Firstly, by virtue of the conditions under which they believe themselves and each other to adopt beliefs and goals in dialogue. The important inclusion to the benevolent approach according to which a lone belief regarding the speaking agent's goal is sufficient, is the receiving agent's preference. Autonomous agents can also

change their existing beliefs and goals as a result of dialogue action. Secondly, agents' dialogue actions are not assumed to always be sincere and open expressions of their mental states. They convey the mental state they wish to induce in the hearer, and that they have a goal for this to be recognised, but this may or may not conform to their own attitudes with respect to the propositional content of the conveyed information. Every dialogue action is a veracious or mendacious, concealing or revealing expression of the speaker's mental states. In conclusion, agents have control over their own mental states in terms of both what is acquired and what is revealed, but control over the flow of information in dialogue between multi-agents, is shared. Each participant has only partial control over the process. Using dialogue as a means of manipulating the mental states of others and achieving desired postures in such a context, therefore requires the nature of dialogue to be strategic.

Dialogue is assumed as strategic interaction between autonomous agents. Strategically rational dialogue action is performed relative to the speaker's belief that the goal of inducing a particular mental state in the hearer will be achieved, and that any subsequent dialogue action on the part of the hearer be in the speaker's favour. In other words, the goal state and receiver's expected subsequent dialogue action are in the speaker's interests. As well as enabling dialogue actions to be strategic tools with which the speaker may achieve her goals, this provides a basis for the hearer to reason about the speaker's goal. It assists in interpretation, especially of non-serious utterances such as lies and irony. It also provides two alternative strategic objectives according to which speakers may commit themselves to particular goals.

CHAPTER 7 : Testing and evaluating the framework

7.1 Introduction

The details of the proposed strategic basis for cooperative linguistic action between multi-agents have now been elaborated in the preceding chapters. The properties of agents and multi-agents, in particular those related to cooperative interaction in the light of conflicts between autonomous agents, have been stated and formally represented. The remaining task is to examine these as an inferential basis for reasoning about speech actions. Do they in fact offer the benefits over previous frameworks claimed in earlier chapters, and which have motivated this research programme? Are they sufficient to enable agents to potentially resolve conflict situations? Can cooperation then emerge between autonomous agents, without imposed benevolence, or assumptions about sincerity?

The aims of this chapter are firstly to demonstrate and test this theoretical framework for modelling dialogue, and secondly to evaluate it in terms of the stated research objectives. Some indication of the historical development of ideas, different methodologies and approaches is also given.

7.2 The methodology

There are two means of testing the framework in order to provide answers to the questions posed above. One way would be to create multi-agents with the requisite properties in the form of computer programs, and analyse their interactions over test conflict situations. The alternative is to take existing dialogues with elements of conflict and/or deception and so on, and analyse them retrospectively; does the theory explain the phenomena which exist here? Is it therefore predictive with respect to such dialogue phenomena? Linguists regularly use this latter approach for the testing their theories, and for which there are many standard conversational exchanges used as examples:

A: "Do you want some coffee?

B: "Coffee would keep me awake", is one such standard example of the use of implicature,

A: "That was a lovely meal", is another of irony, and

A: "The audience snored throughout the film", is another of metaphor.

This research has also adopted the latter strategy, and detailed analyses of two dialogues are given in sections 7.4 and 7.5. The reasons for taking this approach as opposed to the programming option are as follows:

As explained in chapter 4, the focus of this research is the development of a computational theory in Marr's terms, or a specification. This means that it is the nature of the problem of modelling cooperative dialogue and its solution which are of prime interest, and not the determination of specific mechanisms for implementing such a solution. An analysis of examples which illustrate the problem in terms of the proposed theoretical solution is therefore an appropriate means of testing and evaluating the theory. Success in attempting to physically recreate the phenomena of which the examples are comprised on the other hand, would inevitably be dependent upon particular features of the mechanisms employed. These are of course an important research issue, but for a later stage once the content of the theory itself has been established. Writing programs then can helpfully assist in the detection of syntactic bugs. This chapter's use of examples is a much more appropriate methodology to the initial determination of theoretical content, as opposed to correct theoretical syntax.

In addition and as described in chapter 4, building upon and extending existing research, requires the employment of a similar methodology. This is inclusive of the same means of testing and evaluation in order that useful comparison can be made with that previous research. This research builds on the work of Cohen and Levesque (1987b), who tested and evaluated their theory via examples of requests, direct and indirect, serious and non-serious, such as: "Wash the floor", "Get the hammer", or "Go jump in the lake". Their theory explains/ predicts these dialogue phenomena in terms of mental states, their relation to action, and properties of cooperative agents, and without computational implementation. In fact, the implementation of interpreters for any of the extended logics such as epistemic and temporal logics, are still research issues in their own right (Reichgelt, forthcoming book). Moore (1980) and Appelt (1982, 1985) used reified approaches whereby the modal theory is translated into first order predicate calculus and

consequently there were suitable automated theorem provers in existence, and more recently there have been some attempts at building theorem provers for the direct implementation of modal logics (Jackson & Reichgelt, 1987). However, the option of testing a theoretical framework such as this one by writing programs to act as agents in a negotiation scenario still awaits either more developments in the implementation of epistemic logics, or an alternative approach altogether to the modelling of agents.

7.3 Example dialogues and historical approaches

Primarily the examples examined during the course of the research programme were all instances of situations involving conflict between agents. Two whose analyses are described in detail in this thesis in sections 7.4 and 7.5, are as follows: The first is a negotiation between representatives of a union of electricians and their management, in which dialogue is used to resolve an existing conflict. The second is an extract from "Othello" which was selected in order to examine more closely the use of deception and strategic reasoning.

The examples needed to be such that they would offer insights into the complexities of "real life" multi-agent conflicts and interaction. These include agents having multiple goals, reasoning with beliefs as opposed to knowledge, having control over the flow of information between them in terms of being autonomous in belief and goal adoption, as well as having choices where truthfulness and openness are concerned. The electrician's union/ management negotiation is such an example of a "real life" interaction. It is a complete and real negotiation, as opposed to being an experimentally generated context.

At one stage early in the research programme, protocols of an experimental negotiation were used for analysis. The context was that of the game the "Battle of the Sexes", in which the two players both want to spend the evening together, but one wants to go to the ballet and the other to boxing. In this case, the two players were students whose disputed chosen venues for the evening's entertainment were a rock concert and a classical concert. Interesting strategies and attempted manipulations of each other's mental states were evident, but the fact of it being an experimental situation caused some problems. The players were actors role-playing, and therefore without true commitment to the goals of the interaction,

and yet with commitment to their own goals, such as hoping to please the experimenter, or being amusing. This context was used initially with a very different theory of dialogue than that of Cohen and Levesque (1987b), and consequently also that which is described by this thesis. Incorporating ideas from Reichman (1981, 1985), and Birnbaum (1982), Flowers (1982), McGuire, Birnbaum & Flowers (1981), the experimental conversation was analysed according to structural elements known as "argument procedures". Examples of these were: present-challenge, acknowledge-challenge, be-defensive, present-inducement, ...and many more. Each of these was subcategorised according to the goals of a particular interaction, associated with which were rules of inference, which were influenced by the work of R. Cohen (Cohen, 1980). Another taxonomic approach briefly attempted was the characterisation of the actions as types of "strategic moves", distinguished by factors such as the source of the conflict, the directness of the move, the type of deception employed, and so on. These investigations and this experimental context were abandoned as understandings were gained of the constraints of taxonomies and structural approaches as argued in Chapter 1, and correspondingly, the benefits of considering action as determined according to elementary principles of agenthood. The electrician's union negotiation finally also provided a solid "real life" context and useful set of example dialogue actions, for such an analysis.

The second example dialogue in section 7.5, is an extract from the play "Othello". This is not a "real life" interaction. It is a play, but one which represents a "real-life" situation, and this being of great interactive complexity. This does not make the theoretical insights gleaned from its analysis potentially any less valid in the real world. In fact, the analysis of a play is advantaged by focussing on the significant aspects of interactions, without extraneous deviations. Other such instances of the use of literarature are provided firstly by Howard (1971) who used "The Caretaker" by H. Pinter, to analyse his ideas concerning metarationality. He offers a metagame analysis of the relationships between the three central characters. Contexts from Shakesperian plays such as "Measure for Measure" have previously been used in analysis of game-theoretic issues (Schelling, 1960, Colman, 1982). Perrault (1987) also refers to an element contained within the plot of "Othello" to illustrate one aspect of his analysis. I found that the issues of love, death, jealousy, and treachery in plays such as "Othello", contrast well with the relatively "flat" issues which concern what are regularly thought of as computable applications. They provide an extremity and concentration of the issues. The analysis of Iago's relationship with Othello, and the dialogue in the play within which he uses lies and concealment of both beliefs and

goals, resulted in a much greater understanding of strategic reasoning than I might ever have appreciated if I'd restricted myself to more orthodox examples.

This philosophy behind the use of plays as appropriate contexts for theoretical testing and evaluation, has some similarities to that behind the AI methodology of using microworlds. These are a different means of similarly focussing on particular issues. However, microworlds are contexts such as "blocks world", in which the focussing occurs as a result of drastically constraining and limiting the context in the hope that the resulting insights will be applicable in more realistic domains. This has been found to often not be the case. SHRDLU (Winograd, 1972) for example, is a program which simulates the operation of a robot arm manipulating toy blocks on a table and which maintains an interactive dialogue with the user. According to Wilks, SHRLDU's power in problem-solving comes from its employment in such a limited and simple domain. It is unlikely that its methods would be appropriate if extended to a larger domain (Wilks, 1974). Plays such as "Othello" should not suffer similarly as tools for analysis. The complexities of the context are not removed in order to focus on particular issues. All the elements are there; it is their extreme quality and/or relative emphasis which is different from "real life", and which assists focussed analysis.

7.4 The electricians' negotiation

Appendix 1 comprises a complete unedited transcript of a real trade union versus management negotiation (Morley & Stephenson, 1977). It is analysed here as an example of a context in which cooperation and stability emerges from agents interacting with some goals in common, and others in opposition. The agents use dialogue as a means of actually resolving or removing the conflicts between them. They are autonomous; there is no imposed benevolence or sincerity.

The negotiation is analysed by tracing the course of the dialogue, and examining individual utterances and exchanges. The aim is to investigate whether the theoretical principles concerning cooperative multi-agent interaction proposed in this thesis as the basis upon which rational agents generate and interpret speech actions, offer explanations for the speech actions which exist. Those principles which are of primary interest to this research, are summarised in the next section, 7.4.1. The analysis in

section 7.4.2 focuses on these. In 7.4.3 the relevance of this theoretical framework to the evolution of cooperation in the example, is discussed.

7.4.1 A summary of the principles of cooperative multi- agency under scrutiny

1. Agents believe themselves and others to be able to effect postural changes in the multi-agent system, by effecting changes to the beliefs and goals of other agents. The postures of conflict, cooperation and indifference are defined according to configurations of mental states. Representations of these allow agents to recognise existing postures and determine the means of achieving desired postures, according to the model of agenthood .

2. Agents believe themselves and others to be able to potentially manipulate desired changes to other's mental states, and therefore the posture of the system, using dialogue actions. This is conditional upon either evidence in the world or the preferences of the receiving agent. The conditions for adopting a mental state on recognition of it as the goal according to which another agent has generated a dialogue action, are represented as a property of agenthood. Included is the possibility of adopting another's mental state which is contradictory to one in existence. The notion of agent autonomy over the acquisition of information in dialogue is therefore embodied in this property.

3. Negotiation is a type of dialogue whereby multi-agents act on the basis of committed goals to secure agreement on a matter of common concern, and over which disagreement currently exists. Negotiation is a form of strategic interaction, for which agents require a strategic rationality. This means that autonomous, rational agents choose dialogue actions on the basis of firstly, the mental states they want to induce in the hearer, secondly whether they believe this mental state will be achieved, and thirdly subjective expectations of subsequent action on the part of the hearer also being in their interests in this way.

4. Agents understand themselves and others to be able to interact strategically. Adoption of a mental

state on recognition of it as the goal according to which another agent has generated a dialogue action is also therefore dependent upon the hearer's assessment of the speaker's consideration of it as strategically rational.

5. Agents determine strategies according to an understanding of the principles described in 2. and 3., and the nature of interests as defined.

6. Agents believe themselves and others to have control over the information they reveal. They believe themselves and others to be able to choose to perform dialogue actions which are mendacious and/or concealing expressions of a mental state. However, agents are assumed to use revealing and veracious expressions unless the receiving agent has any beliefs which may suggest to the contrary. The nature of veracious, mendacious, concealing and revealing expressions are defined. Agents are assumed to have beliefs concerning the implications of the use of these expressions, which play a role in the determination of preferences as to their use.

7.4.2 The electricians' informal negotiation - utterance analysis

The negotiation was between three electricians and their management representatives. The complete transcript from Morley & Stephenson (1977), can be found in Appendix 1. The cause of the dispute was that the management of the factory required the electricians to be available for callout in case of breakdowns, on bank holidays as well as other days of the year. The electricians did not want to be committed to this; it meant they could never go out or away with their families. Initially they were determined to refuse all offers of money or time off in lieu. The management were equally determined that bank holidays had to be covered, and covered by their own electricians. The conflict was finally resolved with an agreement that two extra men be engaged, and each of the five cover only one bank holiday a year which is rotated between them.

The first two pages of exchanges between the union, collectively represented as U, and management , collectively represented as M, concern the gathering of information. The exact position of each with

respect to the proposition c, is being established.

c = U-to-be-available-for-callout-on--bank-holidays

The conflict situation which exists between them by the end of this stage in the dialogue i.e. at the bottom of page 231 of the transcript, can be represented as:

(G-CONFL-M M U c) and (G-CONFL-M U M c).

For example:

(G-CONFL-M M U c) = (BMB M U ((GOAL M ◊c) ∧ (GOAL U ◊~c)))

∧ (P-R-GOAL M (GOAL U ◊c) q) ∧ (P-R-GOAL U (GOAL M ◊~c) r)

This says that it is mutually believed between M and U that they have a difference in goal related to c. Both of them have a P-R-GOAL to eventually change the other's goal, only to be abandoned if this is achieved, becomes impossible to achieve or the reason for this goal is no longer true. q is M's reason for a commitment to the goal of U adopting M's goal; r is U's reason for a commitment to the goal of M adopting U's goal. Both M and U have representations of this conflict, but although they may or may not have beliefs as to what the other's reasons are, nothing in the conversation has as yet clarified for U what q is, or for M what r is.

From an understanding of the nature of persistent goals (C&L Def: 10), both understand the bases upon which such goals can be dropped. Both want the other to drop their current persistent goals. Their plan for future purposive actions should therefore be aimed at either making the other believe the current goal has become impossible to achieve, or the reason for wanting it is no longer valid. M make the first move in this direction. They go for the latter option, but need some clarification first as to the nature of r. The subgoal which the following action is intended to satisfy is the forming of a belief on the part of M, regarding r. Is r a belief of U's that U should be earning more money?

1. M1: ..Is this merely an attempt on your part to negotiate some price for this external to the agreement...?

This action is a request for U to respond by performing an action from which the subgoal mentioned above can be satisfied. M's plan must be that if the subgoal is satisfied, and if satisfied such that M then has a belief that r is U's belief that U should be earning more money, M has a potential means of making U believe not r (i.e. offering more money), following which U's persistent goal should be dropped.

U recognises M's goal in performing the request as M wanting U to perform a speech action from which M can form a belief about r and its relation to money. U adopts this goal of M's, according to Assumption 6:

$$(BEL\ U\ (GOAL\ M\ \Diamond(DONE\ U\ a))) \wedge (PREFER\ U\ (DONE\ U\ a)\ \sim(DONE\ U\ a)) \supset (P\text{-}R\text{-}GOAL\ U\ (DONE\ U\ a)\ (BEL\ U\ (Gd\text{-}STRAT\ M\ U\ \Diamond(DONE\ U\ a))))$$

2. U1: No, we do not. We won't accept. We don't want to do it in any case. We just want bank holidays as bank holidays. We simply want the time off. We don't, we won't accept it. .....We don't want money for it. We don't want money for it. .....

M responds as follows:

3. M1: How would you suggest then that we deal with this now, as a company? I mean, you're part of the community in this respect.

Again, M is making a request for U to perform some speech action from which M can form a belief. This time the required belief is what U might believe q to be. What will U offer from which M should believe not q?

4. 5. and 6. U2: ..I just wondered if you could have a certain person.....It seems ridiculous if you've got, the plant has got to close down...Now you've got to have coverage....I just wondered if you could have a certain person, say, on these holidays that you could say if there is something happening one Saturday or on a bank holiday you get in touch with him. And try and get one of the electricians. Instead of one electrician being on call, perhaps..

In other words, U want M to form a belief that U are not the only ones who could do the work. There could be some "certain person" to do it instead. U must believe q to be a belief on the part of M that U are the only ones to do the required work. M recognises U's goal that M believe not q. M currently believes q. According to Assumption 4, M will only change his belief to that desired by U if in the current circumstances, M prefers not q to q. M currently does not prefer not q to q. At first, he does not concentrate on this sub-conflict, (B-CONFL-I U M q). This is represented as:

(BEL U ~q) ∧ (BEL U (BEL M q))

∧ (P-R-GOAL U (BEL M ~q) (G-CONFL-2 U M c))

This says that U believes there is a difference in belief with respect to q between M and U, and U has a persistent goal to change M's belief with respect to q, relative to the existence of the goal conflict with respect to c. In other words, if there is no longer the conflict between them about U being available for callout on bank holidays, the goal for M to believe not q is dropped.

M could have responded to this by attempting to make U drop their belief not q, and believe q. They do this later, but first M continues to attempt to establish and then alter U's belief r. Perhaps r is a belief that U have a right to bank holidays as free time. To persuade them that not r:

> 7.8. and 9. M1: Yes, but let's, now let's get back to the case in point. We've got so far, Bill, to the point where they've had this document from John, which was about three months after I got here....I say bank holidays were never brought up by you...never brought up by you, nor was I aware of them.

M's goal that U is to recognise, is that U form a belief that M believes U has no right to demand free bank holidays. This is not stated directly. According to this theory of speech actions as described in chapter 1, the inferences from which communicated goals can be recognised are based upon an understanding of the various principles of rational agenthood and cooperative multi-agent interaction. The latter include properties concerning postures, and thus inference about M's goal in making the

actions 7.8. and 9. are made within the context of the dispute, for which each side believes they and the other have a representation i.e. the pattern of mental states described by (G-CONFL-M M U c) and (G-CONFL-M U M c). M must also believe U to understand the relation between this belief and their goal that eventually not c. U does not respond to this challenge. Since M continues by saying that the work needs to be covered, and what does U suggest, U takes the opportunity to continue pursuing the conflict over q:

10. and 11. U2: Well, as I've just said, I'm not disagreeing with you. You have one man that you can contact, say, on management, if....Well you're bound to get somebody.

Now M responds to this challenge regarding q:

12. . M1: No, no no. Let's sort of be practical about this. Let's suppose now.

At this point U informs M of what in fact, r is:

13. 14. and 15. U1: The idea of not wanting to work is so that we can go out. That's the idea... That's the one day you get er a holiday, then you've... got to sit at home.

M now knows that r is a desire by U to be able to go out on bank holidays, as well as a belief that this is inconsistent with c. M also knows better now, how to try to make r false. He must either make U believe that they don't want to be able to go out on bank holidays, or make this desire consistent with c. The following utterances are an attempt at both:

16. and 17. M1: Now, in the past. We run sweepstakes, for example. A variety of things are in fact going on...None of you have been called out over Christmas.

U still prefer to believe r than not r. They do not adopt M's recognised goal for them to change their belief to not r:

18. U: We're not going anywhere.

At this point, both sides have established what the others reasons are for the conflict with respect to c. In a straightforward way, each agents preferences have been tested by the other in the current context. For either side to now change the other's mind, the context needs altering. In other words, the agents need to be assessing their preferences in the light of some new information. There is the following exchange:

19. M1: Now, this leaves us holding the baby in fact. We have in fact possibly to do something. But how are we to do this? How in fact are we to cover this? Can you suggest to me some way out? Or, in fact, are we saying, "Well that's your bloody problem?" (Very long pause) And if you feel that, then, let's say so...

20. U1: Yes we. We can say it, but we still come back to the same em....

21. 22. and 23. M1: No. No. No. Because it seems to me I can't make you come in. But we have a problem in fact to cover certain eventualities, and it seems to me that we, we will not be able to do so. For...the very reason that we can't call on you....for any reason.

Here M is introducing a new factor. M's strategic reasoning for the generation of the above actions would be as follows:

M has the goal that U form a belief that M and the factory which M represents, is being put in a difficult spot by U. M believes that U will recognise this goal. He believes that U forming such a belief is in his interests. It is a subgoal to the goal of U believing not r. It is based upon a belief that if U adopts this belief, then in further exchanges, U will be assessing his preferences about being free to go out on bank holidays and the incompatibility of this with being on call to work, but in the context of this new belief. M's belief that U also believes there is a mutual belief between them that U and M have to

continue to work together in future, is an important factor in M's considerations about whether this strategy will work or not.

Secondly, M believes U will adopt the desired belief. The basis for this is that U should either acknowledge its "truth", or prefer it. It is unlikely that U will simply prefer to believe that they are the "guilty" party in jeopardising the factory, but perhaps M and U have always dealt fairly and honestly with each other before. From Assumption 8, U can believe that M believes what he's saying. If M believes it, then this will affect their future interactions. U should also believe M to have performed the action according to it being a good strategy for him. Finally, M expects U's subsequent actions to be in his favour. This is also the case. Any action is better than the current deadlock, and at least this introduction of a "guilt" factor, moves the goal posts. U makes no indication at this stage, whether the belief is adopted or not. Later in the interaction at 34., 35., and 36. however, he does.

At this point, the negotiation continues with U performing actions whose strategies are aimed at either M believing not q, which is the belief that that there are other workers other than U to do the work, or reasserting r. r is U's desire to be able to go out on bank holidays and the inconsistency of this with being available for work. Alternatively, M should eventually believe that his goal is impossible:

24. U1: No, I don't want anything out of it. I just don't want to do it. ..I don't want money for it. I don't want a day off in lieu for it. I just don't want to do it, bank holidays.

M continues with his "guilt" strategy. At the same time he offers suggestions of taking on additional workers and different rotas which are unacceptable to U. These suggestions being generated according to a strategic rationality would mean that M believes U will recognise but not adopt his goals. Therefore, the subsequent actions on the part of U are in M's favour, because they will be refusing his apparent compromises. For example:

25. M1: Now suppose we recruit an electrician, and say, 'Well now, part of your job will be in fact to cover these days.'

26. U1: What, bank holidays? Oh, the guys wouldn't think that's fair.

27.M1: We'll ask the recruit.

28. U1: No, no, no, no. For us to sit at home, go, to go out there five days and leave Jo Soap in for work.

29.and 30. M1: Well how can we do this then?....You don't want to do it and you don't want anyone else to do it!

31. U1: No I think its totally unfair that anybody should be asked to do the complete fill at bank holidays.

32. M2: ...You don't want to share and you don't want anybody to do it at all.

If M does not in fact, have a goal to recruit a new electrician, then these actions from which U should recognise the goal to recruit, are mendacious expressions of his mental state. They are strategic in that they are according to his real goal which he also expects U to adopt and therefore U's subsequent actions to be negative to the apparent goal, but satisfying the real goal. Also, if U recognises this and does not believe the apparent goal to be M's good strategy, then since the utterance must have been made according to M's assessment of it as a good strategy, then there is some other goal U is expected to satisfy.

M's plan is that this in conjunction with the "guilt" strategy, will eventually mean that U have no choice but to back down:

33. M1: ...just puts us back to square one, doesn't it? ...Er by creating a situation in which you couldn't resist the fairness of the situation to get involved again. ...I have no solution. You have circumscribed me to such an extent that I can't find a solution, because not only are you going to say I have a right to determine, but you will not have me create a set of circumstances which in any way makes me feel I ought to help.

Exchanges like these continue further. M refers to "mutual concern" and so on. U indicates weakening by showing that he has in fact adopted the belief that he is responsible for putting M and the factory in difficulty. He apologises at 34., and expresses regret at making a fuss at 35., and admits having a conscience about the problems at 36. Finally therefore, M makes the suggestion which was eventually accepted by U at 37.

U's acceptance, and abandonment of their goal for not c as well as the P-R-GOAL for M to adopt this as M's goal, was the case because U eventually was persuaded by M that not r. The conditions were created where U preferred to believe that U's desire to be able to go out on bank holidays and still be available for callout were in fact compatible. M then also abandoned their P-R-GOAL:

(P-R-GOAL M (GOAL U c) q) because (GOAL U c) had been achieved. Thus conflict between M and U and U and M with respect to c, no longer existed. The conditions of the definition were no longer satisfied.

7.4.3 The electricians' informal negotiation - conclusions

Section7.4.2 comprised a retrospective analysis of a dialogue where cooperation and stability emerged in a multi-agent system without imposed benevolence or sincerity conditions. Postures were represented according to the definitions provided in chapter 5. These were shown to provide the contexts within which desired mental states were recognised and ascertained. Utterances were shown to have been generated purposively and strategically, according to the conditions for belief and goal adoption in dialogue as strategic interaction laid down in chapter 6.

In addition, cooperation was seen to have emerged from the conflict, whilst each side was operating according to both individual and collective rationalities. New patterns of behaviour were established; this flexibility allowed the system to survive and cooperative stability to evolve appropriately to the existing conditions. If either side had cooperated with the other without such a negotiation however, but merely because agreement was imposed upon them, the conditions which created the conflict, such as the beliefs and desires concerning holiday working from either perspective, would not have been dissipated. The rigidity of behaviour caused by such subordination can simply result in the total breakdown of the

collectivity of the system at some future occasion requiring interaction for their joint benefit.

The example supports the ideas proposed concerning the positive role of the expression and resolution of conflict in the form of negotiations, to the maintenance and evolution of cooperation - especially in contexts where the agents will interact again in the future. Explanations at the level of individual utterances and exchanges, for the behaviour of the agents during this negotiation and by which the cooperation was achieved, have been given in terms of the properties of cooperative multi-agent interaction proposed.

7.5 "Othello", Act III, Scene III.

Iago's reasoning in the generation of certain utterances from Act III, Scene III of "Othello" by Shakespeare has been analysed according to the proposed theoretical framework. This example was selected as a context in which the speaker uses deception both in the form of concealed intentions, and misrepresentation of the true nature of his mental states.

It was chosen in order to focus on this issue of deception in strategic interaction, especially in a context where the consequences of discovery of the deception are potentially hazardous. Othello's reactions and the probability of success in deception are therefore crucial components of Iago's reasoning prior to the performance of both mendacious and concealing acts. The example is extreme in that in most "real-life" and human contexts, agents do not believe themselves or others to be quite as manipulative and devious as Iago! The consequences of deception are also rarely quite so severe. However, in its extremity, this example proved very influential in the development of the theory of strategic reasoning. It was also crucial to understanding the role and nature of preference. Some insights into this developmental process are given in section 7.5.3.

This is also an example in which conflict is generated by the speaker as a means of achieving his goal. This is a reversal of the emphasis of this thesis, which concerns the use of dialogue in the resolution of conflicts. However, if the theory is correct, then agents should be able to use their knowledge of postures and the power of dialogue, to create any posture.

The role of the postural representations to the directing of change in the multi-agent system is quite

vividly demonstrated by this example. In the previous section comprising the electrician's union negotiation example, both the union and management had representations of postural relations between them which reflected a conflict with respect to the same proposition. On the basis of these, both were engaging in the dialogue as a means of establishing certain of the other agent's beliefs and preferences including the issues that commitment to their contradictory goals were relative to, as well as attempting then to create the conditions whereby the goals would be dropped. In contrast, this example comprises a context whereby only Iago has a representation of the particular conflict between himself and Othello. As a result of Iago's skillful deception, Othello has no such representation. He can therefore neither interpret Iago's actions as strategic attempts to manipulate his mental states towards a particular goal, nor has he any basis for attempting to direct alterations in Iago's mental states to his own advantage.

7.5.1 The context

The relevant section of Act III, Scene III, can be found in Appendix 2. Othello[1] is an army general and Iago his close friend and confidante, as well as Othello's ancient. At the very beginning of the play, Iago expresses his anger at Othello's decision to appoint another man, Cassio, as Othello's lieutenant; a position Iago feels should rightfully have been his. In addition, he feels betrayed by Othello's recent marriage to Desdemona, which usurps the exclusivity of their previous closeness.

In ACT I, Scene III, Iago hatches a plot for revenge on both Othello and Cassio. This involves suggestions of infidelity on the part of Desdemona:

"...I hate the Moor".....

"Cassio's a proper man, let me see now,

To get this place and to make up my will,

A double knavery ...how, how?....let me see,

After some time, to abuse Othello's ear,

That he is too familiar with his wife. " .....

[1] The analysis here is according to just one of several possible interpretations of the play. There is no intention to get involved in literary debates.

Iago's top level goal is revenge. The subgoal to this is to generate conflicts between Othello and Desdemona, and between Othello and Cassio. A means of achieving this, is for Othello to first believe that Desdemona and Cassio are having an affair.

(I = Iago, O = Othello, C = Cassio)

The desired conflicts are:

(B-CONFL-I D O affair) = (BEL D ~affair) ∧ (BEL D (BEL O affair)) ∧

(P-R-GOAL D (BEL O ~affair) (GOAL D remain-wife))

and

(B-CONFL-I C O affair) = (BEL C ~affair) ∧ (BEL C (BEL O affair)) ∧

(P-R-GOAL C (BEL O ~affair) (GOAL C remain-lieutenant))

These are conflicts whereby Desdemona and Cassio believe they are not having an affair, but believe Othello to believe that they are. They each have persistent goals to change Othello's mind, relative to their goals to retain their positions as wife and lieutenant respectively.

A component of Iago's plan is a belief about the power relations between Othello and both Desdemona and Cassio. Also beliefs regarding general attitudes towards affairs, and Othello's insecurities and angers in particular. In other words, Iago must believe that if he can generate these conflicts, neither Desdemona or Cassio are likely to dissuade Othello from banishing and/or punishing them; if this subgoal is successful, then his goal of revenge is most likely to be achieved. The first stage of this plan is to generate the belief in Othello that Desdemona and Cassio are having an affair : (GOAL I ◊(BEL O affair)). Achieving this goal involves mendacious and concealing acts. The concealment is firstly of the mendacity itself, whereby Iago believes that Desdemona and Cassio are not having an affair, whilst performing acts from which Othello should recognise Iago's goal for him to believe the opposite. Secondly, there is also concealment of Iago's top level goal of revenge.

Iago's first problem, however, is that although Iago might have the goal (GOAL I ◊(BEL O affair)) he believes that Othello's existing beliefs include (BEL O ~affair). He also knows that Othello is passionately in love with Desdemona, and therefore he can presume that he will not want to believe she is having an affair: (BEL I (GOAL O ◊~(BEL O affair))). Iago's representation of this conflict can be described as follows:

(G-CONFL-1 I O affair) = (GOAL I ◊(BEL O affair)) ∧
(BEL I (GOAL O ◊~(BEL O affair))) ∧
(P-R-GOAL I (GOAL O ◊(BEL O affair)) (GOAL I revenge))

This says that Iago has a goal for Othello to believe that there is an affair going on, and believes that Othello has a goal to not to believe this. Iago has a persistent goal for Othello to also have the goal to believe it, relative to Iago's goal for revenge.

So, Iago's persistent goal is to induce a goal in Othello to believe that Desdemona and Cassio are having an affair. He understands the basis by which goals can be dropped and new goals generated. He understands the relations between goals and preferences, and preferences and context. Currently, Othello's preferences are such that he prefers to believe Desdemona faithful than not; Iago's subgoal to getting Othello to want to believe Desdemona unfaithful with Cassio, must therefore be to alter the context within which Othello's preferences are operating. Such a goal is attainable on the basis of predictions regarding Othello's beliefs. For example, Iago has several existing beliefs about Othello which are useful to this end:

(BEL I (PREFER O not-cuckolded have-Desdemona))

(BEL I (PREFER O O-have certainty-and-control-over-events
~ O-have certainty-and-control-over-events))

(BEL I (BEL O I-is-honest)), and others, such as believing Othello to be unsuspicious and trusting. Iago refers to Othello in relation to his plan: "as tenderly be led by the nose as Asses are." (Ridley 1958). These beliefs can be used to alter the context. For example, he can perform utterances on the basis of goals for Othello to believe he is possibly being cuckolded and that there is uncertainty over

Desdemona. He knows that Othello is aware that she has already deceived her father in the short time he has known her. He can reaffirm his own honesty and well known reputation as such. The strategic actions he performs are based upon his expectations of Othello's reactions, which in turn are based upon predictions concerning Othello's assessments of himself and others. They concentrate on emphasising those aspects of Othello's beliefs which he reasons make his various goals achievable. He aims at affecting Othello's mental states so that he can no longer hold beliefs as to Iago's honesty and Desdemona's fidelity as consistent beliefs. He must create the conditions in Othello's mind whereby Othello can only believe one of these, and it should be the belief of Iago's honesty which Othello will eventually hold most strongly, and therefore prefer.

Iago will only successfully alter the context for Othello to reassess his preferences concerning the belief about Desdemona's fidelity however, if Iago simultaneously successfully conceals the true nature of his own belief with respect to "affair": (GOAL I ◊~(BEL O (BEL I ~affair)))

The following are a few key utterances and exchanges in the dialogue aimed at the alteration of Othello's mental states as suggested above. These mental states when achieved, form the context within which Othello eventually examines his preferences with respect to the proposition "affair":

1. *Iago.* Did Michael Cassio, when you woo'd my lady,
Know of your love?
2. *Oth.* He did, from first to last:.....why dost thou ask?
3. *Iago.* But for a satisfaction of my thought.
No further harm.

4. *Oth.* What dost thou think?
5. *Iago.* Think, my lord?
6. *Oth.* Think, my lord? By heaven, he echoes me,
As if there were some hideous monster in his thought.....

7. *Iago.* My lord, you know I love you.

8. *Iago.* ....Utter my thoughts? Why, say they are vile and false:.....

9. *Iago.* O, beware jealousy;

It is the green-ey'd monster, which doth mock

That meat it feeds on.....

10. *Iago.* She did deceive her father, marrying you;

And when she seem'd to shake and fear your looks,

She lov'd them most.

11. *Iago.* ...I humbly do beseech you of your pardon,

For too much loving you.

12. *Oth.* ...I do not think but Desdemona's honest.

13. *Iago.* Long live she so, and long live you to think so!

7.5.2 Othello - utterance analysis

The strategic reasoning whereby Iago generates intentions to perform actions such as those above for the purpose outlined above, can be demonstrated with an example:

At the beginning of the conversation between Iago and Othello in Act III, Scene III, Iago says:

1. *Iago.* Did Michael Cassio, when you woo'd my lady,

Know of your love?

The persistent goal for this action to happen next, or in other words the intention to perform this act, would have been generated relative to a persistent goal for Othello to believe:

(BEL O (GOAL I (BEL O there-is- possibly- some-relation-between- Cassio- and-Desdemona -and-Othello))).

This would be relative to it being a good strategy for Iago; it is in his interests, in being a goal he believes will be achieved. Any subsequent action of Othello's is also going to be in his interests. Is this the case?

Firstly, in determining this as a potential goal, Iago must have considered the potential effects of inducing a belief in another, which one does not believe oneself. For example, from Assumption 1:

(BEL I (◊(DONE I lie) ∨ ◊~(DONE I lie))) ∧ (PREFER I (DONE I lie) ~(DONE I lie)) ⊃ (GOAL I (LATER (DONE I lie)))

Iago is faced with two possibilities which are to lie or not to lie about Desdemona and Cassio. His preferences are determined according to maximal consistency with other beliefs and goals at their varying strength, in the current context. Presumably Iago's beliefs about lying - the morality of it, the consequences of it - are not as strong as his goal for revenge.

Secondly he needs to reason as to whether Othello will adopt the belief or not. According to Assumption 6, a belief is adopted on recognition of another's goal for it to be believed, if the agent prefers to believe it than not, and believes it the goal was considered a good strategy by the speaker. In this case, Iago believes Othello is likely to prefer to believe it than not, and will have considered Iago to have considered it a good strategy, simply because he believes Othello's beliefs regarding Iago's honesty and faithfulness to him, are very strong. These will easily affect preferences with regard to what Iago wants Othello to believe, where there is no real challenge or resistance as yet to adopting them.

The third condition for determining good strategy, is whether expected subsequent action on the part of Othello, relative to the successful achievement of the desired belief state, will also be in Iago's interests. Given the beliefs Iago holds about Othello, as suggested in the previous section, such as Othello being trusting and unsuspicious, curious in the face of uncertainties, and believing Iago to be an honest and trustworthy friend, Iago may predict that the subsequent action is likely to be a request for

more information . What might Iago know that Othello doesn't? This is obviously in Iago's interests. However, not enough has been said for any real harm to be done.; whatever the response, it will be favourable to Iago continuing to "plant seeds" in Othello's mind. Othello's reply is in fact:

2. *Oth.* He did, from first to last:.....why dost thou ask?

Iago's subsequent few actions are based on similar reasoning with the desired effect of confirming and strengthening Othello's belief that there is possibly some relation between Cassio, and Desdemona and Othello. For example, in reply to the question above:

3. *Iago.* But for a satisfaction of my thought.
No further harm.

The persistent goal for this action to happen next would have been generated relative to the persistent goal for Othello to believe the following:

(BEL O (GOAL I (BEL O I-has-some-thoughts -relating-Cassio-and-Desdemona -and-Othello-and-the-word-"harm"-is-applicable -to-these))).

There is little to be gained from further detailed analyses of this and others of Iago's utterances in the excerpt. They would look much the same as the example already provided. The reason is that in this example, as opposed to the previous electrician's union example, the dialogue is devoted to the one goal of altering Othello's mental states such that he is susceptible to taking on the goal of believing Desdemona unfaithful. For this, Iago already has several beliefs regarding relevant beliefs of Othello. These were outlined in section 7.5.1. Iago is not using the dialogue to attempt to establish these. There are also no little sub-conflicts as there were in the previous example.

The P-R-GOAL for Othello to have a goal to believe the affair, is eventually successfully achieved and made obvious to Iago at the point in the scene where Othello says:

14. *Oth.* Villain, be sure thou prove my love a whore,

Be sure of it, give me the ocular proof,

Or by the worth of man's eternal soul,

Thou hadst been better have been born a dog,

Than answer my wak'd wrath.

The conflict (G-CONFL-1 I O affair) then no longer exists; Othello has adopted Iago's goal. Othello has therefore become receptive to actually adopting the belief that Desdemona has been unfaithful, and Iago then concludes this element of his overall plan for revenge by introducing evidence in the form of a handkerchief. However, the entire plan does in fact fail later in the play when Othello kills both Desdemona and himself, and the plot is disclosed. Cassio gets further promotion and Iago faces trial and torture.

7.5.3. "Othello" - an alternative analysis

During the development of this strategic framework for cooperative dialogue, analysis of Iago's reasoning about the use of the deception described in the previous sections, was very influential. It gave rise to the ideas related to the importance of the receiving agent's expected response. In the other examples analysed, such as an experimental "Battle of the Sexes" context described briefly in section 7.3, if another agent chose not to adopt the speaker's communicated belief or goal, then not too much harm had been done. A plan had failed. Perhaps the speaker tries again differently, or gives up. In just a few instances however, it can be crucially important that the speaker believes that a component of a plan will be successful before implementing it. Some instances of deception come into this category, and this is one of them. If Othello chose not to believe Iago, then Othello would have to reason that Iago was not a loyal, honest and trustworthy friend, and in fact his action was one of treason. This must be an important element of Iago's reasoning in the generation of his plan.

Metagame analysis (Howard, 1971) was described in section 3.2.5.1. It concerns a choice of action dependent upon predictions of the other agent's action. The evident connection with the ideas of

Schelling (1960) and Goffman (1970) concerning strategic moves and mutual assessments prompted me to do a metagame analysis on Iago's choices of action plan early in the research programme. This will now be described:

The possible outcomes resulting from Iago lying and thereby attempting to induce a belief in Othello that Desdemona had been unfaithful to him are assigned a relative numerical preference, or utility according to Iago's goals and believed goals of Othello, as follows:

(i) Desdemona and Cassio may be punished by Othello for their supposed infidelities. Here Othello suffers the loss of a wife and colleague, and Iago becomes very important as Othellos' only trustworthy friend - all Iagos' goals are satisfied (Iago 4 . Othello 1)

(ii) No action on the part of Iago, and thus also Othello, allows Iago to retain his position with Othello, although achieving no more than maintained love and trust(Iago 3 . Othello 4)

(iii) No action by Othello in response to Iago making the lie also achieves no revenge, but the possibility of mistrust makes the outcome less preferable for both parties(Iago 2 . Othello 3) , and

(iv) the worst outcome for Iago is where Iago is punished as a liar and traitor, and Othello loses a friend (Iago 1 . Othello 2).

Iago's rational choice of strategy is arrived at using Howard's metagame analysis (1971):

(D = Desdemona, C = Cassio, I = Iago, x = implausible outcome)

Basic Game:

	Iago	L (Lie)	N (Not Lie)
	A (Punish D and C)	4,1	x
Othello	B (Punish D)	1,2	x
	C (No action)	2,3	3,4

First level Metagame:

Given that Othello can only act after Iago, and that it is infeasible that Othello would punish either Desdemona and Cassio, or Iago, if no lie was made, then Othello only has three plausible metastrategies (theoretically there are 3^2). i.e.: A.C (Punish D and C if lie is made, else no action), B.C (Punish Iago if lie is made, else no action), and C.C (No action regardless of Iagos' action).

	Othello	A.C	B.C	C.C
Iago	L	4,1	1,2	2,3
	N	3,4	3,4	3,4

Second level Metagame:

Iago now has 2^3 metastrategies eg. LLL (lie regardless of Othellos predicted action), LNN(lie if Othello expected to choose A.C, otherwise choose not to lie.) and so on.....

	Othello	A.C	B.C	C.C	
	LLL	4,1	1,2	2,3	
	NNN	3,4	3,4	3,4	
	LNN	4,1	3,4	3,4	*
	LNL	4,1	3,4	2,3	
Iago	NNL	3,4	3,4	2,3	
	NLL	3,4	1,2	2,3	
	NLN	3,4	1,2	3,4	
	LLN	4,1	1,2	3,4	

Iagos' dominating metastrategy is LNN . This means he should lie if Othello is expected to choose A.C. This is the move where Desdemona and Cassio are punished if the lie is made and Othello becomes more dependent on Iago, but Iago should not make the lie and abandon the current plan if

Othello is expected to make any other of his possible moves. However, it can be seen on the matrix, that this is only the dominating metastrategy if Othello does not know it is Iagos' choice. Otherwise Othello could maximise his own payoff by selecting the B.C or C.C options. Thus Iago must adopt his metastrategy of LNN, but at the same time attempt to convince Othello that his chosen metastrategy is NNN which is one where he does not lie regardless, within which Othellos' payoffs are maximal. This is an example of the strategic use of misrepresentation of intentions (Richelson 1979, Howard 1971, Colman 1982).

Choosing which action to actually execute from the selected metastrategy can only be done on a prediction of the other agents action. Thus Iagos' beliefs about Othellos' attitudes, and principles concerning the interaction of attitudes, are now the required elements for reasoning:
Iago believes that Othello wants Desdemona as a wife, and Othello wants Cassio as a colleague. He can thus infer, that Othello wants to believe that Desdemona and Cassio love him, are honourable and trustworthy, and therefore will do him no harm. Iago also believes that Othello believes that Iago is honourable and trustworthy, and thus will not cause him harm. Therefore, once presented with the proposition that Desdemona has been unfaithful to Othello with Cassio, Othello can either:

(i) believe the lie and infer that his original beliefs about Desdemona and Cassio are incorrect, or
(ii) not believe the lie and infer that his original beliefs about Iago are incorrect, or
(iii) not believe the lie, but assume Iago to have been misguided, or mistaken, and thus allow all beliefs about Desdemona, Cassio and Iago to remain the same.

By examining the payoff matrices, assuming that Othello would have a similar ordering of preferred outcomes in his own model, Iago can assume that Othello would be aware that believing the lie and punishing Desdemona and Cassio implies inevitable personal suffering, and so a preferred outcome can be achieved by not believing the lie which is option (ii) or (iii). But, acting against Iago as a liar and traitor also leads to a less than optimal outcome for Othello. Thus, on the basis of these outcomes alone, (iii) is Othellos' most rational choice of action. If Iago reasoned that Othello's action would really be (iii), then there is no point in him continuing with his plan. It is obvious from a deeper analysis of the play, that Iago knows he can in fact, manipulate Othello to the point of actually wanting

to believe Desdemona's infidelity. Iago must have beliefs about Othello which, in conjunction with an understanding of the nature of beliefs and goals which this analysis has not captured, might lead him to come to a different conclusion. For example, Iago believes, he has a reputation for honesty; he is frequently referred to as "honest Iago". Othello has not known Desdemona for long; Iago believes that Othello passionately loves her, but has not had time to know to trust her. Iago believes that Othello is easily influenced, cannot cope well with uncertainties, and is very insecure about his deficiencies in the world of society (related to his colour, and Desdemona and Cassio both being high class) (Ridley M.R. 1958). In other words, **Iago also knows that some of Othello's beliefs are held much more strongly than others.** The payoff relations are therefore not as straightforward as indicated above.

This understanding lead to the work of Rokeach (Rokeach, 1975) and Quine (1970), and to a greater understanding of the stratification of beliefs and goals, as described in chapter 2. This was crucial to the development of the notion of preference, and the theory given in chapter 6 concerning the control and autonomy an agent has over the adoption of beliefs and goals in dialogue.

7.5.4 "Othello" - conclusions

A metagame analysis for determining best action whatever the other agent might do next, is potentially useful as a means of examining the consequences of all possible plans. In conjunction with an understanding of dialogue as a manipulator of beliefs and goals and the basis upon which this occurs, a sub-plan to increase the likelihood of the desired outcome could then be generated from predictions concerning the other agent's beliefs, goals and preferences. However, as a model for choices of individual dialogue action, metagames seem inappropriate in requiring foresight of all possible retorts and calculating payoff relations for each one. The strategic rationality which is suggested alternatively as the basis for dialogue action in contexts involving conflict, operates according to expectations that any subsequent action will be favourable to the speaker once a particular mental state is successfully induced, and this is determined according to assessments of the other's preferences, beliefs and goals. In other words, these assessments limit the alternatives to be considered.

The role and interest of the metagame analysis of the "Othello" example, is not so much in its results per se, but as a causal influence to theoretical development. It helped to confirm various intuitions, such as those concerning the power of dialogue as a manipulator of mental states, and the role and nature of preference. The example also contrasts well in terms of the necessity for this level of analysis with other conflict situations which do not involve deception.

The analysis in the earlier sections of 7.5.1. and 7.5.2, demonstrated the role of existing and desired postures, represented according to the definitions provided in chapter 5. Iago's representation of a conflict posture between himself and Othello provided the basis for his strategic use of dialogue actions as a means of altering Othello's mental states, in order to therefore alter the context in which Othello's future actions were made, such that these would be in his favour. The orientation of these strategic actions, and the expectations of success were shown to be based upon Iago's assessments of Othello's beliefs, goals, preferences regarding himself, Iago, Desdemona and Cassio, in context. Othello however, was responding to Iago on the basis of a different postural representation. This was as a result of Iago's deception regarding his true beliefs and goals, and Othello's existing but incorrect beliefs regarding Iago's honesty and love for him. Iago's actions were analysed and explained according to the properties of strategic multi-agent interaction given in chapter 6, concerning autonomy over the acquisition of mental states and control over information revealed.

7.6 Conclusions

This research was described in the introduction as being a theoretical preliminary to future implementations of multi-agent cooperative systems for use in DAI or HCI, which can use dialogue to negotiate and resolve differences. In this way such systems can relate flexibly in an unpredictable and changing world such that cooperative action can be maintained, or even evolve. This chapter has described the testing and evaluation of the theory developed towards this end. It was applied to two dialogues comprising firstly a record of a real interaction, and secondly a literary scenario. Both of these were selected as particular examples which demonstrate certain phenomena, and the theory was then used to explain those phenomena. The examples incorporate negotiation and deception, and are explained

according to a theory of strategic reasoning in dialogue as a means of manipulating different postures between autonomous agents.

It is important to acknowledge that there are types and aspects of multi-agent interaction which this theory does not explain. The examples were selected as those which would test specifically the issues for which the theory was intended. Even so, not all the elements of these issues have been tackled. As mentioned at the end of section 3.2.5.2 for example, a useful facility for a system modelling multi-agent dialogue in many day-to-day situations, would be the ability to discern the situations that require the computational effort of a full strategic analysis from the situations that don't. It is not as simple as saying that as long as there is some conflict, reasoning strategically is always appropriate. This is actually an example of a much more general issue which is relevant in many areas of artificial intelligence. There are invariably a number of ways of approaching a problem, and some of these are more computationally expensive than others. The more general the method, the more variation it can handle, but at cost. Considering all dialogue as strategic interaction is a general approach, and the theory presented in this thesis, but it requires a lengthy and expensive reasoning process. Recognition of those specific instances where actions can be made with less processing and without necessarily detrimental consequences would result in more efficient systems.

There are undoubtedly also other aspects of the psychology of negotiation and cooperative dialogue which this theory of strategic interaction does not cover, and for which further study is necessary before practical application can be envisaged. However, this chapter has demonstrated a framework which even as it stands, caters for contexts which previous frameworks were neither designed nor concerned with. This as another step towards the future computational modelling of dialogue in cooperative systems, is suggested as one which offers greater potential for wide and flexible "real world" application.

CONCLUDING REMARKS

This thesis describes theoretical research into the nature of cooperative multi-agent dialogue interaction. Its relevance is as a contextual basis within which future computational agents can reason about dialogue action. Such agents would be components of systems designed according to linguistic principles whereby speech actions are interpreted and generated on the basis of an understanding of what it is to rationally and cooperatively interact. The specific aim has been to uncover principles of cooperative, rational multi-agent interaction, which acknowledge the positive role of conflict to cooperation, and enable the negotiation of differences without imposed benevolence or sincerity. These principles have been explicitly and formally represented in this thesis, in terms of the mental states of interacting agents and the relations between these and action. They have been tested as a basis for dialogue action and evaluated as a means of manipulating postural change.

The motivation for the research is described briefly in the introduction, and in more detail in chapter 3, as a concern that existing AI research into cooperative systems takes an unrealistic view of cooperation; one that does not allow the necessary flexibility of action for survival in an unpredictable and real world. The prevailing view whereby agents take on other's goals simply because they have them, is considered benevolence rather than cooperation. Social psychology studies into the nature of cooperation described in chapter 5, indicate that human conflicts play a crucial and positive role in the maintenance and evolution of cooperation in social systems. The expression of conflict demands a reevaluation of norms of behaviour and conditions, for example. It enables appropriate adaptation in changing conditions. It is suggested that mechanical agents engaging in the kinds of cooperative tasks for which they are currently being envisaged, such as planning, construction, teaching, and advising for example, will be subject to the same real world conditions as their human counterparts. Attempts to program conflicts out of automated multi-agent systems by suggesting they need not exist, avoiding them, or imposing solutions from supervisor processes, can only therefore be practicable in an entirely predictable and controlled environment, but not the real world. The results of other research also discussed in chapter 5, indicate that cooperation evolves amongst self-interested agents, and there is no need for imposed benevolence (Axelrod, 1984).

The linguistic theory adopted from Cohen and Levesque (1987b) whereby communication is

considered as grounded in a theory of cooperative, rational interaction as outlined above, is described at the end of chapter 1. Chapters 2 and 3 together describe the properties of single and multi-agents which make up this strategic theory of cooperative multi-agent interaction, the details and formal expression of which unfold through chapters 4, 5 and 6. The major points of the theory can be very briefly summarised as follows:

Conflict, cooperation and indifference are defined as alternative postures which describe a relation between an agent with respect to another agent and a proposition. They are characterised and correspondingly recognisable as alternative patterns of mental states. Since dialogue is a means of potentially altering agents' mental states, agents therefore have a means of potentially altering the postural relation between them. However, if agents believe themselves and others to not be benevolent, then there is no assurance that a dialogue action will result in the speaker's desired posture. A hearer recognising the speaker's goal that she adopt a particular mental state, may adopt this goal as her own and the desired mental state result, or she may not. In other words, agents believe themselves and others to share control over the outcome of their dialogue actions. Agents are autonomous over what they acquire in dialogue. In order to potentially achieve a desired mental state in another, agents therefore need to understand the conditions under which beliefs and goals are adopted in dialogue. In the absence of evidence of truth in the world, beliefs and also goals are adopted according to the agent's preferences. Existing beliefs and goals can be dropped on the same basis; there is no requirement that new information only be acquired in dialogue if it is not in contradiction with the mental states with exist. This is crucial to agents' understanding that conflicts can be resolved, and differs from previous frameworks unconcerned with conflict as an issue in cooperative interaction. Agents can therefore be induced to "change their minds". Beliefs and goals are adopted also relative to a belief about the basis according to which the speaker became committed to them, this being a notion of good strategy. Strategically rational agents act according to goals which they believe will be attained, and to which they believe any response on the part of the other agent will also be in their favour. Agents generate and interpret dialogue actions on the basis of believing themselves and others to generate dialogue actions according to a strategic rationality. Dialogue is strategic interaction. Finally, there is no assumption that agents are truthful or sincere. Dialogue actions can be veracious or mendacious, concealing or revealing expressions of a mental state of the speaker.

In chapter 7, selected dialogue actions from a real trade union negotiation between an electrician's union and their management are explained according to the above theory. Also in chapter 7 is an analysis of an extract from "Othello" which focusses on the use of deception in strategic interaction. The latter example was very influential to theoretical development, but the prime motivation for focussing on conflict in multi-agent systems was to generate a framework in which the positive aspects of this could be realised. This thesis therefore presents a framework for negotiation; a framework for dialogue as a means of resolving conflicts.

The theory presented in this thesis offers a wider notion of cooperative multi-agent interaction than previous AI research. It incorporates insights gained from other related fields of study, such as social psychology, philosophy and game theory. The nature and role of conflict to the evolution and maintenance of cooperative systems, the nature of dialogue as strategic interaction, the nature of agents' control over the flow of information between them in dialogue - these, together with arguments as to their relevance to computational applications, are the major contributions of this work. In addition, and related to this, the research described here has pointed out how much work is still to be done. Preferences, for example, emerged as an important property of agents, reflecting circumstantially-based relationships between different beliefs and goals. These relate directly to the strengths with which the different beliefs and goals are held which support the contending issues. It is not a trivial task to represent beliefs and goals at varying relative strengths, and to account for constant changes in this during the course of an interaction. This issue has consequently been avoided in research to date; agents are modelled with one belief or goal as equally as important and hence retainable or rejectable as the next. Other representational problems were discussed in chapter 4 and also chapter 2, particularly with respect to the formal approach which was adopted. The problem of expressing relative probabilities for example, and idealised rationalities. This thesis has not been concerned with representational issues to any greater extent than in selecting the most suitable currently available approach, in order to undertake the research. Obviously however, continuing interest by other researchers into these issues is of great relevance to the potential for both practical and further theoretical developments of this work.

This research has built on the little research there is in AI with regard to the nature of agenthood, and the particular context of agents interacting with each other and creating social or multi-agent systems. As mentioned earlier, elements of the theory were borrowed from social psychology, philosophy and

game theory, which are disciplines with more experience in these matters. I hope this thesis points out the value to AI firstly, of such an interdisciplinary approach, and secondly, the importance of the continued gathering of more and more theoretical insights into the nature of rational agenthood and multi-agency, before the development of HCI and DAI computational applications.

Cooperation was the issue primarily selected here as requiring some attention. Whilst investigating the nature of multi-agent cooperation and the role of conflict in this, other issues became apparent, such as the nature of agent autonomy and control over the flow of information in dialogue, for example. Only the surface of this issue has been touched here however, particularly with respect to the nature and varied uses of deception in dialogue. As pointed out by Goffman (1970) and referred to in chapters 3 and 6, deception is a part of everyday communication. There are many grades and types of dishonesty, and I believe this should not be considered as necessarily a sinister concern. Telling nothing but the whole and complete truth in all circumstances is more than often inappropriate, if not time-wasting, or at worst potentially destructive. Just as the view of conflict as distasteful prevented social psychologists in the past from dealing with it and recognising its positive role in social systems, so similar attitudes to issues of sincerity and openness will generate a blinkered approach to the computational modelling of multi-agent dialogues. This thesis has only briefly raised this as an issue, but I believe it to be one worthy of further discussion and study.

Another issue which has not been dealt with in full here is that of negotiation. In fact, the view which this thesis implicitly presents in its emphasis, of all conflicts being resolvable via dialogue, suffers from the same lack of realism just used to criticise other research in which all agents are assumed benevolent and sincere. Only a brief suggestion was made in section 3.2.3 of further research taking into account the necessity for agents to be able to assess whether and/or when a choice of "best action" in the circumstances is more appropriate. There is a great wealth of research into bargaining and strategic interaction from other disciplines such as social pychology, economics and trade union studies, for example. All of this, and more, warrants further investigation.

BIBLIOGRAPHY

This bibliography includes details of papers and books which proved influential to the development of the research, in addition to those specifically referenced in the text of the thesis. The former are indicated with an asterisk.

Allen, J. and Perrault, C.R. Participating in dialogues: Understanding via plan deduction. *Proceedings of the 2nd National Conference, CanadianSociety for Computational Studies of Intelligence*, Toronto, 1978.

Allen, J.F. A Plan-Based Approach to Speech Act Recognition. TechnicalReport 131, Dept. of Computer Science, Univ. of Toronto, Canada, 1979.

Allen, J. & Perrault, C.R. *Analysing intention in dialogues. Artificial Intelligence* Vol. 15, No 3, pp143-178, 1980.

Allwood, J. Linguistic Communication as Action and Cooperation - study in Pragmatics. Gothenburg Monographs in Linguistics 2, 1976.

*Allwood, J., Andersson, L-G. & Dahl, O. Logic in Linguistics. Cambridge University Press, 1977

Anscombe, G.E.M. *Intention. Cornell University Press, Ithaca*, 1963.

Appelt, D.E. Planning Natural language utterances to satisfy multiple goals. Technical Note 259, SRI International, Menlo Park, CA., 1982.

Appelt, D.E. Planning English Sentences. Cambridge Univ. Press, 1985.

Austin, J. L. How to do things with words. Oxford University Press, 1962.

Axelrood, R. The evolution of cooperation. Basic Books Inc. N.Y., 1984.

Bach, K. Communicative Intentions, Plan Recognition and Pragmatics: Comments on Thomason, Litman and Allen. In Cohen, P. and Levesque, H. (Eds.) Readings for Formal Theories of Communication. 1987 Linguistics Institute, LI127A Vol. 1, Stanford University, California, pp165-182, 1987.

Bach, K. & Harnish, R.M. *Linguistic Communication and Speech Acts.* MIT Press, Cambridge Mass, 1979.

Barwise, J. & Perry, J. Situations and Attitudes. Bradford Books, MIT Press, 1983.

Bennett, P.G. Hypergames: Developing a model of conflict. Futures, pp 489 - 507, December 1980.

Birnbaum, L. Argument Molecules: A Functional Representation of\ Argument Structure. Proceedings of the National Conference in Artificial Intelligence, American Association for AI, University of Pittsburgh, Pennsylvania. pp 63-65, 1982.

Bok, S. Lying. Moral Choice in Public and Private Life. Vintage Books, N.Y., 1978.

Brand, M. Intending and Acting. MIT Press, Cambridge, Mass., 1984.

Bratman, M. Intention, Plans and Practical Reason. Harvard Univ. Press, 1987.

Brehmer, B. Cognitive Factors in Interpersonal Conflict. In Druckman D. (Ed.) Negotiations: Social-Psychological Perspectives. Sage Publications, Beverley Hills/London, 1977.

*Brickman, P. Social Conflict: Readings in Rule Structures and conflict relationships. D.C.Heath and Co. Lexington Mass., 1974.

*Brown, P. and Levinson, S. Universals in Language useage : politeness phenomena. In Goody, E.N. (Ed.) Questions and Politeness. Cambridge University Press, 1978.

*Bruce, B. Robot plans and Human plans: Implications for models of Communication. Report 314, Bolt Beranek and Newman, Cambridge Mass., 1984.

*Bruce, B. & Newman, D. Interacting Plans. Cognitive Science Vol 2., pp195-233, 1978.

Cacciari, C. Communication Rituals in a doctor-patient Setting; politeness strategies. Paper presented at the first international conference on Applied Psycholinguistics, Barcellona, June 1985.

Castaneda, H.N. Thinking and Doing. Reidel, Dordrecht, Holland, 1975.

*Chellas, B.F. Modal Logic - an introduction. Cambridge Univ. Press, Cambridge, 1980.

Cherniak, C. Minimal Rationality. MIT Press, Cambridge, Mass., 1986.

Clark, H. & Carlson, T.B. Speech acts and Hearer's Beliefs. In Smith N.V. (Ed.) Mutual Knowledge. Academic Press, N.Y., 1982.

Clark, H. & Marshall, C. Definite reference and Mutual Knowledge. Paper presented at the Sloan Workshop on Computational Aspects of Linguistic Structure and Discourse Setting. University of Pennsylvania, 1978.

*Coddington, A. Theories of the Bargaining Process. Allen & Unwin, London, 1968.

Cohen, P. On Knowing What to Say: Planning Speech Acts. Technical Report No. 118. Dept. of Computer Science, University of Toronto, Canada., 1978.

*Cohen, P. & Levesque, H. Speech acts and the recognition of shared plans. Proceedings of the third biennial conference of the Canadian Society for Computational Studies of Intelligence, Victoria B.C., 1980.

Cohen, P. & Levesque, H. Speech acts and rationality. Proceedings of the23rd annual meeting of the Association for Computational Linguistics. Bell Communication Research, Morristown N.J., 1985.

Cohen, P. & Levesque, H. Persistence, Intention and Commitment. Technical Report No. 415, AI Centre, SRI International, California, U.S.A.,m 1987a. Also: Proceedings of the 1986 Workshop at Timberline Lodge, Reasoning about action and plans. Kaufmann Inc., 1987a.

Cohen, P. & Levesque H. Rational Interaction as the basis for Communication. Technical report No. 89, Centre for the Study of Language and Information, Stanford University, California, U.S.A., 1987b.

Cohen, P. & Perrault, C.R. Elements of a plan-based theory of speech acts. Cognitive Science 3, pp177-212., 1979.

Cohen, R. Understanding arguments. Proceedings of the Canadian Society for Computational Studies of Intelligence, pp 272-279, 1980.

Colman, A. Game Theory and Experimental Games - A study of Strategic Interaction. Pergamon Press, 1982.

Cooley, C.H. Social Organization. Scribner's Sons, New York, 1909.

Coser, L. The Function of Social Conflict. Glencoe Ill. Free Press, 1956.

Coser, L. Social Conflict and the Theory of Social Change. The British Journal Of Sociology, Vol. 8, pp 197-207, 1957.

Coser, L. Continuities in the Study of Social Conflict. The Free Press, New York, 1967.

Davidson, D. Actions, Reasons and Causes. Journal of Philosophy, Vol. 60, pp 685 - 700, 1963.

*Davis, L.H. Prisoners, Paradox and Rationality. American Philosophical Quarterly, Vol. 14 No. 4, pp 319-327, 1977.

*Davis, R. & Smith, P.G. Frameworks for Cooperation in Distributed Problem Solving. Proceedings of IEEE Transactions for Man, Systems and Cybernetics, Vol. SMC-11, No. 1, pp 61-70, 1981.

Davis, R. & Smith, P.G. Negotiation as a Metaphor for Distributed Problem Solving. Artificial Intelligence, Vol. 20, pp 63-109, 1983.

Dawkins, R. The Selfish Gene. Oxford University Press, Oxford, 1976.

Dennett, D.C. Brainstorms. MIT Press, Cambridge, Mass., 1978.

*Deutsch, M. The Effect of Motivational Orientation upon Trust and Suspicion. Human Relations, Vol. 13, pp 123-139, 1960.

Deutsch, M. Conflict and its Resolution. In Smith, C.G. (Ed.) Conflict Resolution: Contributions of the Behavioural Sciences. University of Notre Dame Press, London, 1971.

*Deutsch, M. The Resolution of Conflict - Constructive and Destructive Processes. Yale Univ Press, New Haven London, 1973.

*Donnellan, K. Putting Humpty Dumpty together again. Philosophical Review, Vol. 76, pp 203-215, 1968.

*Doran, J. The computational Approach to Knowledge Communication and structure in Multi-actor systems. In (Eds.) Gilbert, G.N. & Heath, C. Social Action and Artificial Intelligence. Gower, Aldershot, 1985.

Doran, J. Distributed Artificial Intelligence and the Modelling of Sociocultural Systems. Computer Science Memorandum, CSM-87, University of Essex, 1987a.

Doran, J. Multi-actor systems and the Emergence of Organisations. Proceedings of the 7th Alvey Planning SIG, Martlesham, Ipswich, 1987b.

Dretske, F.I. Knowledge and the Flow of Information. Blackwell, Oxford, 1981.

Druckman, D. Conflict of Interest and Value Dissensus: Two Perspectives. In Druckman, D. (Ed.) Negotiations: Social-Psychological Perspectives. Sage Publications, Beverley Hills/London, 1977.

Dubin, R. Power and Union-Management Relations. Journal of Conflict Resolution, Vol. 1, Adm. Sci. Quarterly No. 2, pp 60-81, 1957.

Durfee, E.H., Lesser, V.R. & Corkill, D.D. Cooperation through Communication in a Distributed Problem Solving Network. In Huhns M.N. (Ed.) Distributed Artificial Intelligence. Pitman, London, 1987.

*Dyer, M.G. Affect Processing for Narratives. Proceedings of Proceedings of the National Conference in Artificial Intelligence, American Association for AI, pp 265-268, 1982.

Edmondson, W. Spoken discourse: A Model for Analysis. Longman, London, 1981.

Edwards, W. Behavioural Decision Theory. In Edwards ,W. & Tversky, A. (Eds.) Decision Making. Penguin, Middx., 1967.

*Elsom-Cook, M. Towards a framework for human-computer discourse. In Johnson, P. & Cook, S. (Eds.) People and Computers: Designing the interface. Cambridge Univ Press, 1985.

Engel, P. Functionalism, Belief and content. In Torrance, S. (Ed.) The Mind and the Machine - philosophical aspects of artificial intelligence. Ellis Horwood Ltd., Chichester, 1984.

Fagin, R. & Halpern J.Y. Belief, Awareness and Limited Reasoning. Proceedings of the 9th International Joint conference on A.I., CA., pp 491-501, 1985.

Fikes, R.E. & Nilsson, N. STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. Artificial Intelligence, Vol. 2 , pp 189-208, 1971.

Flowers, M. On being Contradictory. Proceedings of the National Conference in Artificial Intelligence, American Association for AI, University of Pittsburgh, Pennsylvania, pp 269-272, 1982.

*Fodor, J.A. Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. In Haugeland, B. (Ed.) Mind Design. MIT Press, Camb., Mass., 1981.

Fox, M.S. An organizational View of Distributed Systems. Proceedings of IEEE Transactions for Man, Systems and Cybernetics, Vol. SMC-11, No. 1, pp 70-80, 1981.

*Galliers, J.R. A Strategic Framework for Multi-Agent Cooperative Dialogue. Proceedings of the Eighth European Conference on Artificial Intelligence, Munich, pp 415-420, August, 1988.

*Galliers, J.R. A Definition of Cooperation (but not Benevolence), for multi-agent planning. Report of the Alvey-IKBS-Expert Systems Research Theme, 7th Planning SIG, Martlesham, Ipswich, 1987.

*Galliers, J.R. A Theoretical Framework for Modelling Conflict and Cooperation in Dialogue. British Psychological Society symposium, London, December 1986.

Gazdar, G. Pragmatics: Implicature, Presupposition and Logical Form. Academic Press, New York, 1979.

Gazdar, G. Speech Act Assignment. In Joshi, A., Webber, B. & Sag, I. (Eds.) Elements in Discourse Understanding. Cambridge University Press, 1981.

Georgeff, M.P. Communication and Interaction in Multi-Agent Planning. Proceedings of the National Conference in Artificial Intelligence, American Association for AI, Washington D.C., pp 125-129, August 1983.

Georgeff, M.P. A Theory of Action for Multi-agent planning. Proceedings of the National Conference in Artificial Intelligence, American Association for Artificial Intelligence, Austin, Texas, pp 121-125, August 1984.

*Georgeff, M.P. Actions, Processes and Causality. Proceedings of the 1986 Workshop at Timberline Lodge. Reasoning about Actions and Plans. Kaufmann Inc., U.S.A., 1986.

Ginsberg, M.L. Decision Procedures. In Huhns M.N. (Ed.) Distributed Artificial Intelligence. Pitman, London, 1987.

Goffman, E. Strategic Interaction. Blackwell, Oxford, 1970.

*Goody, E.N. Towards a theory of questions. In Goody, E.N. (Ed.) Questions and Politeness". Cambridge University Press, 1978.

Gordon, D. & Lakoff, G. Conversational Postulates. In Cole, P. & Morgan, J.L. (Eds.) Syntax and Semantics. Vol. 3. Speech Acts. Academic Press, N.Y., 1975.

Grice, H.P. Meaning. Philosophical Review, Vol. 66, pp.377-388, 1957.

Grice, H.P. Utterer's Meaning and Intentions. Philosophical Review, Vol. 78, No. 2, pp147-177, 1969.

Grice, H.P. Logic and Conversation. In Cole, P. & Morgan, J.L. (Eds.) Syntax and Semantics. Vol. 3. Speech Acts. Academic Press, N.Y., 1975.

*Grice, H.P. Further Notes on Logic and Conversation. In Cole, P. & Morgan, J.L. (Eds.) Syntax and Semantics. Vol. 3. Speech Acts. Academic Press, N.Y., 1978.

Grosz, B. Focusing and Description in natural language Dialogues. In Joshi, A., Webber, B. & Sag, I. (Eds.) Elements in Discourse Understanding. Cambridge University Press, 1981.

Grosz, B. & Sidner C. Attention, Intention and the Structure of Discourse. Computational Linguistics, Vol. 12, Part 3, pp175-204, 1986.

Haack, S. Philosophy of Logics. Cambridge Univ. Press, 1978.

Halpern, J.Y. & McAllister, D.A. Likelihood, Probability and Knowledge. Proceedings of the National Conference in Artificial Intelligence, American Association for Artificial Intelligence, Austin, Texas, pp 137-141, August 1984.

Halpern, J.Y. & Moses, Y.O. Knowledge and Common Knowledge in Distributed Environments. Proceedings of third ACM conference on Principles of Distributed Computing, pp 50-61, 1984.

Halpern, J.Y. & Moses, Y.O. A Guide to the Modal Logics of Knowledge and Belief. Proceedings of the 9th International Joint Conference of AI, Los Angeles, pp 480-490, 1985.

Halpern, J.Y. & Moses, Y.O. Reasoning about Knowledge: An Overview. Proceedings of Theoretical Aspects of Reasoning about Knowledge, pp 1-17, Kaufmann Inc., U.S.A., 1986.

Hare, R.M. Some alleged differences between imperatives and indicatives. Mind, Vol. 76, pp 309-326, 1967.

Harel, D. First-order dynamic logic. Springer-Verlag, New York City, New York, 1979.

Harrison, R. Rational action: Studies in Philosophy and Social Science. Cambridge University Press, 1979.

Haslett, B.J. Communication: Strategic Action in Context. Lawrence Erlbaum Associates, New Jersey, U.S.A., 1987.

Hintikka, J. Knowledge and Belief. Cornell University Press, N. York, 1962.

*Hintikka, J. Logic, Language-games and Information. Oxford Univ. Press, 1973.

Hobbs, J.R. & Evans, D.A. Conversation as Planned Behaviour. Cognitive Science, Vol. 4, pp349-377, 1980.

*Hodges, W. Logic. Penguin Books, Middx., 1977.

*Houghton, G. & Isard, S. Why to speak, what to say and how to say it: modelling language production in discourse. In Morris, P (Ed.) Modelling Cognition. John Wiley & Sons Ltd., 1987.

Howard, N. Paradoxes of Rationality. MIT Press, Cambridge, Mass., 1971.

Howard, N. General metagames: an extension of the metagame concept. In Rapaport, A. (Ed.) Game theory as a theory of conflict resolution. Reidel, Dordrecht, 1974.

*Hughes, G.E. & Cresswell, M.J. An introduction to Modal Logic. Methuen and Co. Ltd., London, 1968.

Huhns, M.N. Distributed Artificial Intelligence. Pitman, London, 1987.

Hume, D. A Treatise of Human Nature. Clarendon Press, Oxford, 1978.

Israel, D. The Role of Logic in Knowledge representation. Proceedings of IEEE Transactions for Computer, pp 37-41, 1983.

*Jackson, P. Reasoning about belief in the context of advice-giving systems. In Bramer,M. (Ed.) Research and Development in Expert Systems. Cambridge Univ. Press, 1985.

*Jackson, P. & Lefrere, P. On the application of Rule-based Techniques to the design of Advice-giving systems. International Journal of Man-Machine Studies, Vol. 20., pp63-86, 1984.

Jackson, P. & Reichgelt, H. A General Proof Method for First-Order Modal Logic. Proceedings of the tenth International Joint Conference on AI, pp 942-944, 1987.

Johnson-Laird, P.N. Reasoning Without Logic. In Myers, T., Brown, K. & McGonigle, B. (Eds.) Reasoning and Discourse Processes. Academic Press, London, 1986.

*Jones, A.J. Game theory : mathematical models of conflict. Ellis Horwood, Chichester, 1980.

*Jones, A.J. Towards a formal theory of communication and speech acts. In Cohen, P. & Levesque, H. (Eds.) Formal theories of communication. Linguistic Institute LI127A, Vol. 1, Stanford University, pp 247-279, 1987.

*Joshi A., Webber, B. & Sag, I. Elements of discourse understanding. Cambridge Univ. Press, 1981.

Kerr, C. Industrial Conflict and its Mediation. American Journal of Sociology LX, pp 230, 1954.

Kidd, A. What Do users ask? - some thoughts on diagnostic advice. Proceedings of Expert systems, pp 9-19, 1985.

Kiss, G. Outlines of a computer model of motivation. Proceedings of the third International Joint Conference on AI, Stanford, pp 446-449, August 1973.

Kiss, G. Agent Architecture. Unpublished HCRL technical report, Open University, Milton Keynes, England, 1986.

Kiss, G. Structures and Computation. HCRL working paper No.2, Open University, Milton Keynes, 1987.

Kolata, G. How can Computers get Common Sense. Science, Vol. 217, pp 1237-1238, 1982.

Konolige, K. A first order formalisation of knowledge and action for a multi-agent planning system. In Hayes, J.E., Michie, D. & Pao, Y. *Machine Intelligence 10*. Ellis Horwood Ltd., Chichester, 1982.

*Konolige, K. Belief and Incompleteness. Report 319, SRI, Menlo Park, CA, 1984.

Konolige, K. A Deduction Model of Belief. Pitman, London, 1986.

*Konolige, K. What Awareness Isn't: A Sentential View of Implicit and Explicit Belief. Proceedings of Theoretical Aspects of Reasoning about Knowledge, pp 241-250, Kaufmann Inc., U.S.A., 1986.

*Korner, S. Practical Reason. Blackwell, U.K., 1974.

*Korner, S. Experience and Conduct: A Philosophical Enquiry into Practical Thinking. Cambridge Univ Press, 1976.

Kornfield, W.A. & Hewitt, C.E. The Scientific Community Metaphor. Proceedings of IEEE Transactions for Man, Systems and Cybernetics, Vol. SMC-11, No. 1, pp 24-33, 1981.

*Kripke, S.A. A Completeness Theorem in Modal Logic. Journal of Symbolic Logic 24, pp 1-15, 1959.

Kripke, S.A. Semantical Analysis of Modal Logic. Zeitschrift fur Mathematische Logik und Grundlagen der Mathematik, Vol. 9 pp 67-96, 1963.

*Kripke, S.A. Semantical Considerations in Modal Logic. Linsky L. (Ed.) Reference and Modality. Oxford Univ Press, 1971.

*Kripke, S.A. Naming and Necessity. Blackwell, Oxford, 1972.

*Kripke, S.A. Outline of a Theory of Truth. Journal of Philosophy, Vol. 72, pp 690-716, 1975.

*Lakemeyer, G. Tractable Meta-reasoning in Propositional Logics of Belief. Proceedings of the tenth International Joint Conference on AI, Milan, 1987.

Lesser, V.R & Corkhill, D.D. Functionally accurate, cooperative, distributed systems. Proceedings of IEEE Transactions for Man, Systems and Cybernetics, Vol. SMC-11, No. 1, pp 81-96, 1983.

Lesser, V.R & Corkhill, D.D. The use of Meta-level control for coordination in a distributed problem solving network. Proceedings of the eighth International Joint Conference on AI, pp 748-756, Karlsruhe, Germany, 1986.

*Levesque, H. A formal treatment of incomplete knowledge bases. Technical Report No. 614, Fairchild AI Lab., Palo Alto, California, U.S.A, 1982.

Levesque, H. A Logic of Implicit and Explicit Belief. Proceedings of the National Conference in Artificial Intelligence, American Association for AI, Austin, Texas, pp 198-202, August 1984.

Levin, J.A. & Moore, J. A. Dialogue Games: Metacommunication structures for natural language interaction. Cognitive Science, Vol. 1, No. 4, pp 395-420, 1977.

Levinson, S.C. Pragmatics. Cambridge Univ. Press, 1983.

Lewis, D.K. Convention: A Philosophical Study. Harvard University Press, Cambridge, Mass., 1969.

Linde, C. Information Structures in discourse. Ph.D Thesis, Colombia University, 1974.

Litman, D. Discourse and Problem Solving. Report TR130, Univ. of Rochester N.Y., 1983.

Luce, R.D. & Raiffa, H. Games and Decisions : Introduction and a Critical Survey. New York, Wiley, 1957.

*Lyons, J. et al. New Horizons in Linguistics. Penguin, London, 1987.

Mack, R.W. & Snyder, R.C. The Analysis of Social Conflict- Toward an Overview and Synthesis. In Smith, C.G. (Ed.) Conflict Resolution: Contributions of the Behavioural Sciences. University of Notre Dame Press, London, 1971.

MacKay, A.F. Professor Grice's theory of meaning. Mind, Vol.81, pp 57-66, 1973.

*Marley, P. Cooperation between Autonomous Robots. Computer Science Memo CSM-19, University of Essex, 1977.

Marr, D. AI; A Personal View./ In Haugeland, B. (Ed.) Mind Design. MIT Press, Camb, Mass., 1981.

*Marr D. Vision. Freeman & Co., U.S.A., 1982.

*Mates, B. Elementary Logic. Oxford Univ. Press, 1965.

Maynard-Smith, J. & Price, G. The logic of Animal Conflicts. Nature, Vol. 246, pp15-18, 1973.

McCarthy, J. & Hayes, P. Some philosophical problems from the standpoint of artificial intelligence. In Mitchie, D. (Ed.) Machine Intelligence, 4 . American Elsevier, pp 463-502, 1969.

*McCarthy, J., Sato, M., Hayashi, T. & Igarashi, S. On the Model Theory of Knowledge. Memo AIM-312, Stanford AI Lab, Stanford, U.S.A., 1978.

McClintock, C.G. Social Motivations in Settings of Outcome Interdependence. In Druckman, D. (Ed.) Negotiations: Social-Psychological Perspectives. Sage Publications, Beverley Hills/London, 1977.

McGuire, R., Birnbaum, L. & Flowers, M. Opportunistic Processing in Arguments. Proceedings of the seventh International Joint Conference on AI, University of British Colombia, Vancouver, Canada, pp 58-60, 1981.

Minsky, M. K-lines: A Theory of memory. In Norman, D. (Ed.) Perspectives on Cognitive Science. Norwood, N.J., Ablex, 1981.

*Mitchell, R. & Thompson, N. Deception: Perspectives on Human and Nonhuman Deceit. State University of New York Press, New York, 1986.

Montague, R. Syntactical Treatments of Modality with corollaries on reflexion principles and finite axiomatizations. Acta Philosophica Fennica, Vol. 16, pp 153-167, 1963.

Moore, R.C. Reasoning about Knowledge and Action. Technical Note No.191, SRI International, California, U.S.A., 1980.

*Moore, R.C. The Role of Logic in Artificial Intelligence. Technical Note 335, SRI, California, U.S.A., 1984.

Morgan, J. Comments on R. Perrault's "An application of Default Logic to Speech Act Theory". In Cohen, P. & Levesque, H. (Eds.) Readings for Formal Theories of Communication, Linguistics Institute, LI127A, Vol. 1, Stanford University, California, 1987.

Morley, I. & Stephenson, G . The social psychology of Bargaining. Allen & Unwin, London, 1977.

Morris, C.W. Foundations of the Theory of Signs. In Neurath, O. et al (Eds.) International Encyclopaedia of Unified Science, Vol. No. 2, University of Chicago Press, 1938.

Perlis, D. Languages with Self-reference 1: Foundations. Artificial Intelligence, Vol. 25 , No. 3, pp 301-322, 1985.

Perrault, C.R. & Allen J.F. A Plan-Based Analysis of Indirect Speech Acts. Proceedings of the American Journal of Computational Linguistics, Vol. 6, No. 3-4, pp167-182, 1980.

Perrault, C.R. An application of Default logic to Speech Act Theory. Report No. CSLI 87-90, CLSI, SRI International, California, U.S.A., 1987.

Pollack, M. Plans as Complex Mental Attitudes. In Cohen, P. & Levesque, H. (Eds.) Readings for Formal Theories of Communication, Linguistics Institute, LI127A, Vol. 1, Stanford University, California, 1987.

Pollack, M., Israel, D. & Bratman, M. Toward an Architecture for Resource-Bounded Agents. CSLI Report No. CSLI-87-104, CSLI, Stanford, California, U.S.A, 1987.

Power, R. The organisation of purposeful dialogues. Linguistics, Vol. 17, pp107-152, 1979.

Power, R. Mutual Intention. Journal for the Theory of Social Behaviour, Vol. 14, pp 85-101, 1984.

Power, R. Joint planning in Conversation. Abstract in the Proceedings of the Structure of Multi-modal Dialogues including voice, Venaco, France, September 1986.

*Power, R. Efficiency in Conversation. Alvey Workshop on Explanation, Surrey University January 8-9, 1987.

*Quine, W.V. Word and Object. MIT Press, Camb, Mass., 1960.

Quine, W.V. The Web of Belief. Random House, N.York, 1970.

Ramsey, A. Knowing That and Knowing What. In Hallam, J. & Mellish, C. (Eds.) Advances in Artificial Intelligence, John Wiley and Sons, Chichester, 1987.

*Rapaport, A. Game theory as a theory of conflict resolution. Reidel Publ. Corp., Dordrecht, 1974.

Reichgelt, H. A Review of McDermott's "Critique of Pure Reason". AI Communications, Vol 0, No. 1, August 1987.

Reichgelt, H. Knowledge Representation: An AI Perspective. Ablex Publishing Corporation, England, forthcoming book.

Reichgelt, H & Shadbolt, N. Epistemic Logic and Cooperative Planning. Submissions to the 7th Alvey Planning SIG workshop, British Telecom, Martlesham, Ipswich, 1987.

Reichman, R. Modelling Informal Debates. Proceedings of the seventh International Joint Conference on AI, University of British Colombia, Vancouver, Canada, pp 19-24, 1981.

Reichman, R. Getting Computers to talk like you and me - discourse context, focus and semantics (an ATN model). MIT press, 1985.

Rich, A. On Lies, secrets and silence; selected prose 1966-1978. Virago Press, London, 1975.

Richelson, J.T. Soviet Strategic Doctrine and Limited Nuclear Operations - A Metagame Analysis. Journal of Conflict Resolution, Vol. 23, No. 2, pp326 - 336, 1979.

Ridley, M.R. (Ed.) "Othello" by Shakespeare W., Methuen, N.Y. and London, 1958.

Rokeach, M. Beliefs Attitudes and Values - A Theory of Organisation and Change. Jossey-Bass, N.Y., 1975.

*Rosenschein, J.S. Synchronisation of multi-agent plans. Proceedings of the National Conference in Artificial Intelligence, American Association for AI, University of Pittsburgh, Pennsylvania, pp 63-65, 1982.

Rosenschein, J.S. Rational Interaction: Cooperation Among Intelligent Agents. Ph.D thesis, Stanford University, California, U.S.A., 1985.

Rosenschein, J.S. & Genesereth M.R. Deals among rational agents. Proceedings of the 9th International Joint Conference of AI, Los Angeles, pp 91-99, 1985.

Rosenschein, S.J. Abstract Theories of Discourse and the Formal Specification of Programs that Converse. In Joshi, A., Webber, B. & Sag, I. (Eds.) Elements of Discourse Understanding. Cambridge University Press, 1981.

Rosenschein, S.J. & Kaebling, L.P. The Synthesis of Digital Machines with Provable Epistemic Properties. Proceedings of Theoretical Aspects of Reasoning about Knowledge, Kaufmann Inc., U.S.A., 1986.

*Rummel, R.J. Understanding Conflict and War, Vol. 3, Sage Publications Inc. Beverley Hills/London, 1977.

Sacerdoti, E. A Structure for Plans and Behaviour. Elsevier North-Holland Inc., Amsterdam, 1977.

*Sadock, J.M. On testing for Conversational Implicature. In Cole, P. Syntax and Semantics. Academic Press, New York, 1978.

*Salmon, W.C. The Foundations of Scientific Inference. University of Pittsburgh Press, 1966.

Sato, M. A Study of Kripke-type models for some Modal Logics by Gentzen's Sequential Method. Ph. D thesis, Research Institute for Mathematical Sciences, Kyoto University, 1976.

Schegloff, E. & Sacks, H. Opening up closings. Semiotica 8, pp289-327, 1973.

Schelling, T.C. The strategy of Conflict. Harvard University Press, Camb., Mass., 1960.

Schlenker, B.R. & Bonoma, T.V. Fun and Games- The Validity of Games for the Study of Conflict. Journal of Conflict Resolution, Vol. 22, No.1, pp 7-38, Sage Publications Inc., 1978.

Schiffer, S.R. Meaning. London. Oxford University Press, 1982.

Searle, J.R. Speech Acts. Cambridge University Press, 1969.

Searle, J.R. Indirect Speech Acts. In Cole, P. and Morgan, J. L. (Eds.) Syntax and Semantics, Vol. 3. Academic Press, 1975.

Searle, J.R. Expression and Meaning; Studies in the Theory of Speech Acts. Camb. Univ. Press, Camb., 1979.

Searle, J.R. Intentionality: An Essay in the Philosophy of Mind. Cambridge University Press, New York City, New York, 1983.

Simmel, G. Conflict. New York. Free Press, 1955.

Shadbolt, N. & Musson, C.L. Cooperative Planning; a foundation for Communicative Negotiation. Proceedings of the 6th Alvey, Planning SIG Workshop, Christs College, Cambridge, 1986.

Shadbolt, N. & Musson, C.L. Cooperative Planning and Cooperative execution. Proceedings of the 7th Alvey, Planning SIG Workshop, Martlesham, Ipswich, 1987.

Sheppard, H. Approaches to Conflict in American Sociology. British Journal of Sociology V, pp 324-342, 1954.

Shoham, Y. Nonmonotonic Logics: Meaning and Utility. Proceedings of the tenth International Joint Conference on AI, pp 388-393, 1987.

Simon, H. Models of Man. John Wiley, New York, 1957.

Simon, H. Administrative behaviour. Macmillan & Co., New York, 1959.

Singer, K. Resolution of Conflict. Social Research VI, pp 230, 1949.

Sloman, A. The Computer Revolution in Philosophy; Philosophy, Science and Models of Mind. Harvester Press, Sussex, 1978.

*Smith, N.V. Mutual Knowledge. Academic Press, London, 1982.

*Sowden, L. That there is a dilemma in the Prisoners Dilemma. Synthese, Vol. 55, pp347-352, 1983.

Sperber, D. & Wilson, D. Mutual Knowledge and Relevance in Theories of Comprehension. In Smith, N.V. (Ed.) Mutual Knowledge. Academic Press, New York, 1982.

Sperber, D. & Wilson, D. Relevance. Blackwell, Oxford, 1987.

Sridharan, N.S. Workshop Report: 1986 Workshop on DAI. AI Magazine, pp75-85, Fall 1987.

Stich, S. From folk Psychology to Cognitive Science: The Case against Belief. MIT Press, Cambridge, Mass., 1983.

Suchman, L. Plans and Situated Actions. The Problem of Human-Machine Communication. Cambridge University Press. 1987.

*Thomas, C.S. *Design and Theory of Metagame* experiments. In Rapaport, A. (Ed.) Game theory as a theory of conflict resolution. Reidel Publ. Corp., Dordrecht, 1974.

Vardi, M.Y. On epistemic Logic and Logical Omniscience. Proceedings of Theoretical Aspects of Reasoning about Knowledge, Kaufmann inc., U.S.A., pp 293-305, 1986.

von Neumann, J. & Morgenstern, O. Theory of Games and Economic Behaviour. Princeton University Press, Princeton, N.J., 1944.

*Walton, R.E. & McKersie R.B. A Behavioural Theory of Labour Negotiations. An Analysis of a Social Interaction System. McGraw-Hill Book Company, New York, 1965.

Wilensky, R. Planning and Understanding. Addison Wesley, Reading Mass., 1984.

*Williams, B. Internal and External Reasons. In Harrison, R. (Ed.) *Rational Action: Studies in* Philosophy and Social Science. Cambridge University Press, 1979.

*Wilkins, D.E. Using Patterns and Plans in Chess. Artificial Intelligence, Vol. 14, pp 165-203, 1980.

*Wilkins, D.E. Domain-independent Planning: Representation and Plan Generation. Artificial Intelligence, Vol. 22, pp 269-301, 1984.

Wilks, Y. Natural Language Understanding Systems within the AI paradigm: A survey and some comparisons. AI memo 237, AI laboratory, Stanford University, 1974.

Wilks, Y. Beliefs, Points of View and Multiple Environments. Cognitive Science, Vol. 7, pp 95-119, 1983.

Wilson, N.L. Grice on Meaning: The ultimate counterexample. Nous, Vol. 4 pp 295-304, 1970.

Winograd, T. Understanding Natural Language. New York, Academic Press, 1982.

Wright, R. Meaning_{nn} and Conversational Implicature. In Cole, P. & Morgan, J.L (Eds.) Syntax and *Semantics*, Vol. 3. Academic Press, 1975.

Ziff, P. On H.P. Grice's account of meaning. Analysis , Vol. 28, pp 1-8, 1967.

APPENDIX NOT COPIED

ON INSTRUCTION FROM

UNIVERSITY