Implementing a Normative Theory of Communication in a Framework for Default Reasoning

by
Andrew Csinger

Technical Report 91-30
November 1991

Department of Computer Science
University of British Columbia
Rm 333 - 6356 Agricultural Road
Vancouver, B.C.
CANADA V6T 1Z2

Abstract

This thesis presents a framework for inter-agent communication, represented and partially implemented with default reasoning. I focus on the limited goal of determining the meaning for a Hearer–agent of an utterance ω by a Speaker–agent, in terms of the beliefs of the interlocutors. This meaning is generally more than just the explicit propositional contents of ω, and more than just the Speaker's goal to convey her belief that ω.

One way of determining this meaning is to let the Hearer take stock of the implicit components of the Speaker's utterances. Among the implicit components of the meaning of ω, I show in particular how to derive certain of its presuppositions with a set of default schemata using a framework for default reasoning.

More information can be extracted from the communications channel between interlocutors by adopting a normative model of inter-agent communication, and using this model to explain or 'make sense' of the Speaker's utterances. I construct such a model expressed in terms of a set of default principles of communication using the same framework for default reasoning.

The task of deriving the meaning of an utterance is similar to the job required of a user-interface, where the user is the Speaker–agent, and the interface itself is the Hearer–agent. The goal of a user-interface as Hearer is to make maximal use of the data moving along the communications channel between user and application.

The result is an integrated theory of normative, inter-agent communications expressed within an ontologically and logically minimal framework. This work demonstrates the development and application of a methodology for the use of default reasoning. The implementation of the theory is also presented, along with a discussion of its applicability to practical user-interfacing. A view emerges of user-modelling as a component of a user-interface.

Contents

List of Tables

List of Figures

Acknowledgements

My parents. They made it possible for me to ask these questions with their love and devotion, and by never telling me to become a doctor.

My wife. My poor, long-suffering Susan. Enough said.

My supervisor. David Poole has been very generous with his time and energy, and bought me that pitcher of Beer when I really needed it.

My advisor. Richard Rosenberg has given me only good advice, and some of his enthusiasm has rubbed off on me.

Michael Horsch. Mike has been instrumental in getting parts of this thesis to make sense, and parts of the implementation to work. He was also there to help me drink that pitcher.

Emanuel Noik. Manny's mission, it seems to me, is to keep me motivated. I don't know from where he gets his boundless energy or vocation. I can only thank him and hope my good fortune continues.

UBC. The university is near some of the most breathtaking scenery in the world. These natural monuments are always around to remind me who I am whenever I start taking myself too seriously.

Chapter 1

Introduction

Common Sense: Those superstitions we learned before the age of eighteen.

—Einstein.

1.1 What this thesis is about

At its highest level, this thesis is about user-interfacing.

My conception of a user-interface is of a support structure for communications between an intelligent agent and an applications program. The user-interface bridges the gap between user and application, forming a channel along which communications can take place. The information-carrying capacity of this channel can be qualitatively described in terms of its bandwidth.

The goal of user-interfacing is to broaden the bandwidth of the communications channel between user and application.

There are potentially many ways to accomplish this broadening. Some that have been suggested are programmable command-decoders, graphical input-output devices, natural-language interfaces, multi-media output, and multi-sensory input-output. I restrict myself in this thesis first of all to interfaces which can be implemented over a conventional serial (teletype-like) channel, and focus further on a natural language style environment. I accomplish the broadening effect by exploiting tacit and implicit components of user utterances, using a theory of communications. I choose to express the additional information gleaned from the utterance in terms of the beliefs of the user-agent. To this end, I build a model of the user based on the utterances she makes. Figure 1.1 is a schema of the domain this thesis is concerned with; this schema is refined in later chapters to show the various sub-components.

A view emerges that a user-modeller can be considered to be a sub-component of the user-interface, and that user-modelling is one of the tasks that a user-interface might be called upon to do in fulfilling its goal of broadening the bandwidth of the communications channel.

Figure 1.1: From Utterance to Belief

At its lower levels, this thesis is about presupposition, about theories of communication, and about implementing these in a default reasoning framework.

1.2 A Theory of Communication

A recurrent theme throughout this thesis is that the communicative content of what is uttered is not restricted to its propositional contents; in addition to what is directly asserted by an utterance, there is a set of propositions which are indirectly implied, and the set of those which are antecedently assumed. Loosely, the first set has been referred to as the implicatures of an utterance, while the latter includes what are known as felicity conditions. I show how to derive a subset of both the implicit and tacit contents of utterances, in terms of the beliefs of the interlocutors involved. Previous work has invariably employed some form of Cooperative Principle, according to which the utterances in a discourse are presumed to adhere to a set of guidelines, itself tacitly represented by the participants in the discourse. I too make use of such principles, but with the desire to capture the realistic departures that are routinely made in the attempt to mislead, to be sarcastic, and so on.

1.2.1 The Implicit and the Tacit

In general, implicatures of an utterance are those propositions which are implied but not directly stated by the utterance. Recent usage, however, has followed the work of Grice[22], who identified certain types of inference which he then named *implicatures;* he further distinguished these into categories with distinct properties. *Conventional* implicatures are those which arise solely from features of the words employed in an utterance, and this thesis is concerned with only this kind of implicature. Henceforth, I use the term *implicature* in this technical sense, and show how some conventional implicatures can be derived from context-situated utterances in the framework of the principles of communication I define.

Tacit phenomena are fundamental to communication. Often expressed in terms of *mutual beliefs,* tacit information is generally held to be known by all members of the group under observation, and further to be known to be known to all these members. In particular, participants in a dialog are usually held to believe that the principles of cooperative communication alluded to above are in effect. In general, elements of world knowledge are also considered to be available to the members of a group, so that this information may go unsaid in conversations among members of a group. This type of tacit information has been referred to as presupposed by an utterance, or by the speaker

making an utterance. I employ the term *presupposition* in its more technical sense, that of the *sentential presuppositions* of an utterance. (See section 4.3). This is a class of pragmatic inference distinguished mainly by its defeasibility in the context of contradictory information, and by its characteristic behavior under negation. I show how sentential presuppositions of varying lexical environments can be derived from context-situated utterances, the cooperative principle, and the implicatures of the utterance.

1.2.2 Representation and Implementation

Tacit phenomena and pragmatic inference are often characterized in terms of their conjectural nature. Defeasibility has long been a distinguishing feature of natural language presupposition, and the maxims of cooperative communication are self-evidently fragile. In Chapter 2, I follow the historical thread of the defeasibility of pragmatic inference from first attempts at formalizing presupposition, to recent work using default reasoning. I see this work as continuing this trend, and the model I present in Chapter 3 is itself completely implemented in a default reasoning framework; I acquire and represent beliefs of agents from their utterances using the **Theorist** [45] framework for common-sense reasoning.

1.3 User-interfacing and User-modelling

Much of my early work was aimed at improving user-interfaces for Computer-Aided Design (CAD) systems, where the efficacy of the interface can be measured qualitatively in terms of the maximum rate of information exchange between user and application [14], [13]; others have recognized this metric and have described it in different terms. In the domain of information retrieval, these words have been written:

> ...improvements in an information storage and retrieval system focus on the idea of improving the cost-effectiveness of the system, in terms of the quality of the information retrieved in relation to the time, effort, and expense of storing and retrieving it.[34]

I have named this qualitative measure of the rate of information exchange between user and application, the *bandwidth* of the communications channel, and sought in the past to increase this bandwidth with a variety of *ad-hoc* techniques suited to the CAD environment.

This thesis pursues and generalizes the idea of expanding the bandwidth of the communications channel between user and application, through the interpositioning of a User–Modeller **UM**. The role of the user-interface expands to include the functions usually attributed to a **UM**. In this thesis I describe how principles of communication along with a related theory of presupposition –both implemented in a framework for default reasoning– constitute a **UM** capable of increasing the bandwidth of the user-application communications channel, in a principled manner. Figure 1.2 is a variation on a conventional view of the user-interface[43]; I have added the **UM** unit to be viewed either as a sub-component of, or as communicating with the user-interface. This schema maps cleanly into the communications domain with the recognition of the user's role as Speaker-agent, and the role of the user-interface as Hearer-agent.

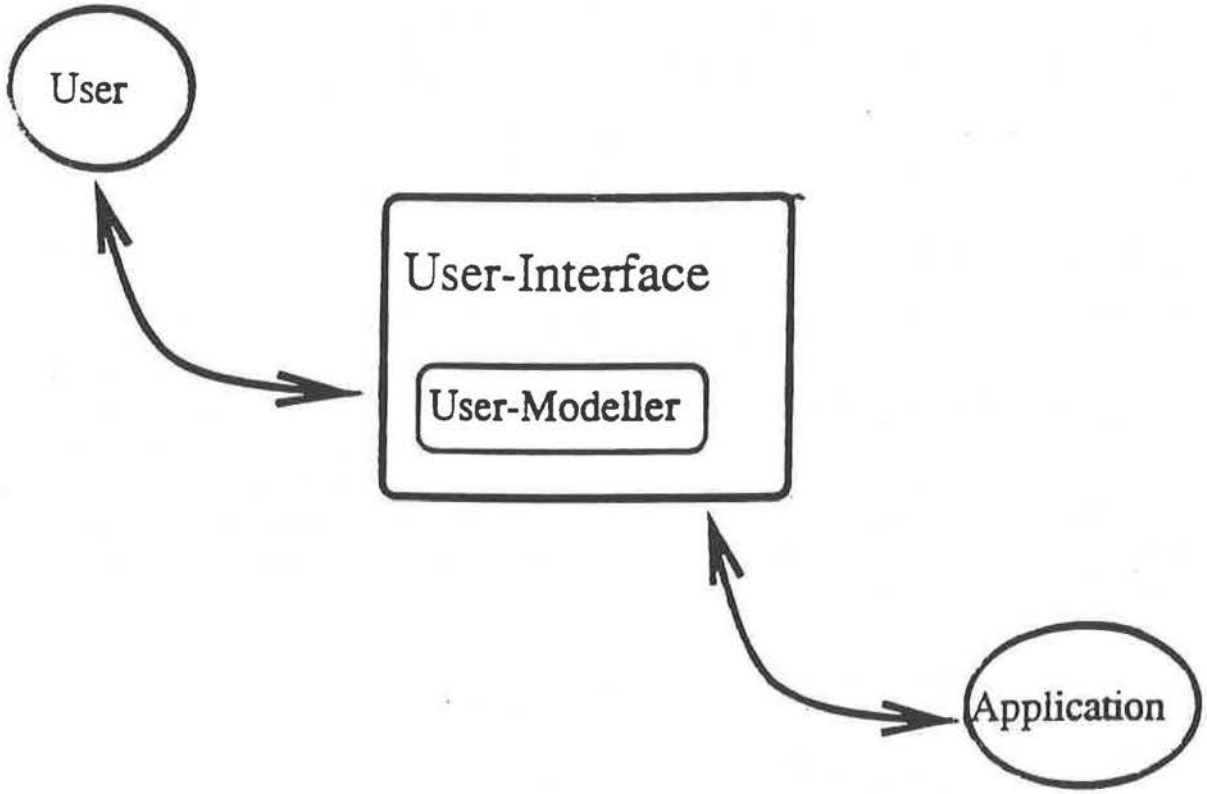

Figure 1.2: The User-Modeller as Part of the Interface

1.4 Priorities

Throughout this thesis, I argue for logical and ontological minimality. I see this work as part of the movement of "minimal AI", which seeks to accomplish its goals with the least posturing about psychological relevance, or cognitive validity.

Certain linguists and psychologists have characterized my position as one of timidity, and have urged me to take a stand on the psychological relevance of the computational architecture set forth in this thesis. I believe they do this from a misunderstanding of the goals of AI in general, and the aims of this work in particular. There is plenty of room for differences in opinion on the former, so I will deal only with the latter objection.

I am not engaged in empirical cognitive science here, but in minimal, empirical artificial intelligence. The approach is minimal, because I try to adopt only those elements of a logical calculus that are *necessary* to accomplish the representational requirements of this study. In particular, I represent (and derive) the following:

- presuppositions of natural language utterances
- principles of natural language communications
- principles for deriving beliefs from other beliefs

I do not argue anywhere that other approaches or representational schemes cannot accomplish the same objectives; I only demonstrate that these objectives can be accomplished in a principled manner within the particular framework for default reasoning that is minimal with respect to its underlying logical calculus. So it may well be that particular connectionist networks and a host of *ad-hoc* logics can implement systems with equivalent characteristics, but I show that these are not *necessary* to achieve the results of this system. I leave it to the psychologists, however, to decide the cognitive relevance of the computational units I describe.

1.5 Organization of this Thesis

In chapter 2, I survey previous work in the areas of presupposition, theories of communication and user-modelling. I explore some of the work done by philosophers and psychologists on belief and rationality, and I introduce the default reasoning formalism which I use to implement my own theory.

Chapter 3 is a consideration of the issues I faced in deciding the eventual path that the implementation would take. Previously unexplored methodological issues are investigated, and some alternative implementation strategies are pursued.

The implementation itself is detailed in chapter 4. Some of this work appeared elsewhere ([17]).

I conclude in chapter 5 with what I consider the contribution of this thesis, along with a consideration of the problems that remain to be solved, and some suggestions that might lead to their resolution.

Chapter 2

Background

To spend too much time in studies is sloth.

—Francis Bacon

In this chapter I trace the lineage of previous work that leads to my research in the pragmatics of communication. There are many dimensions along which a survey of this kind might be made. I pursue the growing recognition in the literature that certain classes of pragmatic inference are *defeasible*, with particular attention to the study of presuppositional phenomena. Early work attempted to stay within the bounds established by classical logic, but these 'semantic' theories appear to be giving way to 'pragmatic' varieties which take into account more than the behavior of truth-functional-connectives in natural language.

Previous work in the formulation of 'cooperative' principles underlying communication is addressed as well. As in the discussion of presupposition, there is a unifying thread of defeasibility running through the literature. This thread has only recently been perceived as indicative of a *default* nature, and I amplify on this point. I discuss the relation of a model of communication to user modelling, and I present some salient issues in previous work. I introduce the terminology with which my own work will be described.

2.1 Presupposition

There are a variety of reasons for studying presuppositional phenomena in natural language, not the least of which is their ubiquity. As alluded to in the motivational preface to this thesis, masterful use of human language involves subtleties which are not captured by even the most detailed analyses of the propositional contents of a discourse.

Linguistic presupposition has been recognized from the start as something peculiarly extra-propositional, a blemish on the uniform face of classical logic. Certainly in the eyes of logicians of the day, the phenomenon had to be accounted for.

Much of the previous literature has been created out of a concern over the 'projection problem' associated with presupposition. This is the study of how the presuppositions of the constituent clauses of a compound sentence 'project' over the sentence. Various perspectives will be considered, and I later argue –following Burton-Roberts and others–

that the concern over projection has been due to previous definitions of the presupposition relation, rather than to the existence of a *problem* with projection as such. Another issue is the behavior under negation of the presupposition relation, and once again, I will consider various attempts to define presupposition in view of this behavior.

2.1.1 History of Presupposition

Despite the movement toward acceptance of the predicate calculus as the language of linguistic representation, it became clear very early in the process that it would place too severe restrictions on expression, and that certain relations manifest in natural language could not be captured with it. Previous study of presuppositional phenomena has typically resorted to various non-standard logics to avoid certain difficulties.

Negation in Natural Language One problem which continues to plague a standard-logic analysis is the following.

Example 2–1:

Sentence 2–1: The king of France is bald.
Sentence 2–2: The king of France is not bald.
Sentence 2–3: There exists a king of France. ■

Example 2–1 is Strawson's [54], although this is in fact a very old story [19]. Both sentences (2–1) and (2–2) are commonly held to presuppose (2–3). The problem arises when (2–3) is false; this is a case of *presupposition failure*. If (2–1) is regarded as false because of the non-existence of the referent, then if the natural language negation is interpreted in the wide-scope sense, (2–2) can only be given the value of true by recourse to the law of the excluded middle. One way out that has been taken is to adopt a tri-valent logic which assigns to (2–1) and (2–2) the third value in the case of presupposition failure [19, 54]. Although this and similar approaches avoid the aforementioned contradiction, they suffer from an inflexibility of application: there are instances where presupposition failure does not deny a truth value from the sentence.

Sentence 2–4: The King of France is (not) a woman.

Sentence 2–4 is intuitively false (true?) in spite of the failure of the presupposition that there is a King of France.

Russell's approach was to represent sentences with presupposed referents as in equation 2.1, which is his proposed logical form for Sentence 2–1.

$$\exists_x(king(x) \wedge \neg\exists_y(y \neq x \wedge king(y)) \wedge bald(x)) \tag{2.1}$$

A natural language negation operator can then be interpreted in various ways: The speaker could be negating the 'kingliness' of the referent, or his baldness or even the existence of the referent. To Russell, natural language negation is thus inherently ambiguous.

Strawson argued for the truth-valuelessness of utterances like 2–1 and 2–2 on the basis of 'pure intuitions' to this effect. Most so-called definitions of 'semantic presupposition' have in fact centered on Strawson's notion, paraphrased by definition 1. The relation Strawson calls necessitation is an implication that does not support *modus tollens.*

> **Definition 1 (Strawson)** *Sentence A* semantically presupposes *sentence B iff sentence A necessitates sentence B, and the denial of sentence A necessitates sentence B.*

From this, Strawson argues, if sentence B is not true, then sentence A is neither true nor false. Thus, semantic presupposition is not classical entailment, because there is no support for contrapositives, and it requires a tri-valent logic. Although there is no pretheoretic or theoretical obstacle to such an account of presupposition, sufficiency is not adequacy in itself, and the semantic approach must stand against the challenges of other theories [38, p81].

It remains for the traditional semantic account to render a mapping of natural language connectives to logical truth-functional connectives, thereby allowing for a purely compositional interpretation of 'projection.'

A *pragmatic* view of presupposition failure is that the utterance is somehow 'infelicitous,' having violated some of the maxims of cooperative communication (see § 2.2.1 and 2.2. Early pragmatic accounts center on proposed solutions to the so-called projection problem, characterized by context-sensitive rules designed to over-ride the normal behavior of purely compositional rules.

The Negation Test and the notion of Defeasibility The following discussion exposes what is called the *negation test,* a criterion of linguistic ancestry which any successful definition of the presupposition relation must accomodate.

It has been argued that both sentence 2–1 and sentence 2–2 presuppose sentence 2–3. This is to say that certain negated lexical environments carry the same presuppositions as their affirmative counterparts. This phenomenon has been promoted as a necessary condition on a relation, for it to be considered presupposition *per se.*

Example 2–2:

> **Sentence 2–5:** The King of France is not bald, because there is no King of France.
>
> **Sentence 2–6:** *The King of France is bald, because there is no King of France. [1] ■

Example 2–2 demonstrates the defeasibility of presupposition. The presupposition of sentence 2–5 is cancelled from within the sentence itself, without upsetting the intuitions of a native speaker. The second clause serves to focus the scope of the negation operator on the existence of the referent, rendering the statement unambiguous. (This is known

[1] Some have argued that sentences of the form *The king of France rules over Normandy, but there is no king of France,* are felicitous in contexts where the first clause refers to the *intension,* and the second clause to the *extension* of *king of France* (referentially opaque and transparent readings, respectively). If the reader's intuitions tend in this direction, I urge that he replace *king of France* in all its occurrences with **present** *king of France.* I am interested in the extensions of the referring terms.

as *internal* negation). The same presupposition of sentence 2–6 cannot be successfully defeated; an *infelicity* results [4]. Along with its behavior under the negation test, the defeasible nature of the presupposition relation is another feature that distinguishes it from other candidate pragmatic inference classes.

Projection The projection into the matrix sentence of the presuppositions of its constituent clauses has been recognized as a problem for theories of presupposition [30],[20].

Example 2–3:

Sentence 2–7: (He stopped singing) and (the audience began to applaud).

Sentence 2–8: He had been singing.
Sentence 2–9: The audience had not been applauding. ■

Horton [27, p78] gives example 2–3 as representative of a class of sentence in which presuppositions of constituent clauses project over the sentence. The presuppositions of the first and second clauses are sentences 2–8 and 2–9, respectively, and both of these project, or become presuppositions of sentence 2–7 itself.

Example 2–4:

Sentence 2–10: My cousin is a bachelor or [my cousin is] a spinster.
Sentence 2–11: My cousin is male.
Sentence 2–12: My cousin is female. ■

Example 2–4 is one in which some of the presuppositions of the clauses do not project over the matrix sentence. Sentences 2–11 and 2–12 are these presuppositions; they are mutually contradictory, and thus do not project. This is an example of cancellation from within the sentence itself. This example is dealt with in more detail in later sections of this thesis.

Karttunen and Peters contributed the "Plugs, Holes and Filters" account of presupposition projection [30]. They divided linguistic environments into three categories, distinguished by their effect on the projection of presuppositions. *Holes* are those environments in which presuppositions always survive embedding, while *plugs* block projection. *Filters* are middle ground, where presuppositions sometimes fail, sometimes project, depending upon *filtering conditions.* There are numerous objections to the approach. First, it is considered unprincipled by some [36], in that the theory grows in complexity when presented with more complex data. The method has been shown to make incorrect predictions [ibid]. And last, is the conflation of the presupposition relation with other pragmatic inference classes. Karttunen and Peters deny the defeasibility of presupposition, thereby losing what I see as its most distinguishing feature. Instead of using defeasibility (via the negation test, perhaps) as a defining characteristic of the relation, they attempt to develop a theory which predicts only those presuppositions that will not subsequently be cancelled.

Gazdar first put presuppositional analysis on a firm pragmatic footing. He argued convincingly in favor of an approach based on consistency, rather than on truth values in projection. He developed a notion of 'satisfiable incrementation,' [20, p131] which, in retrospect, presaged the newer theories associated with default logic and common-sense reasoning.[2] Gazdar recognized and emphasized the defeasibility of the presupposition relation. He proposed rules to generate presuppositions which were to be regarded in some sense as conjectural, and which could be defeated by contradictory presuppositions of clauses in complex sentences, or by inconsistency with context. He called these potential presuppositions, or *pre-suppositions,* emphasizing that they were no more than 'notional entities' that played a 'technical role' in his theory[20]. These pre–suppositions, defined by definition 2, become *actual* presuppositions only if they survive the mechanics of the context incrementation method introduced later by Gazdar.

Gazdar postulates a function for each lexical environment that carries presuppositions, and suggests that the set F of these functions has a cardinality which is "some small finite number," and that

> Obviously one can go further and define f_4, f_5, f_6, etc. for all the other sources of pre-suppositions but, as far as I can see, this is a theoretically trivial task, and I do not propose to pursue it here.

It may be *theoretically* trivial, but it certainly poses a number of difficult practical questions! In particular, the definition explicitly ignores the surface form of the sentences that are its domain, and it remains unclear what the cardinality of F might be. Though perhaps theoretically uninteresting if the set is in fact finite, the actual size will no doubt reflect upon the efficiency of any implementation.

Gazdar proposes, for instance, to capture the presuppositions of a sentence with a factive verb, with the following function:

$$f_1(\phi) = \{\psi : (\psi = \mathbf{K}\chi) \wedge (\phi = X \frown \nu \frown that \frown \chi \frown Y)\}$$

where ν is a factive or semifactive verb, ϕ and χ are sentences, and X and Y are any strings, possibly null. **K** is read as *the speaker knows that.*

Example 2–5:

Sentence 2–13: Oedipus regrets that Jocasta drinks
Sentence 2–14: **K** (Jocasta drinks) ■

Gazdar presents example 2–5, where sentence 2–13 presupposes sentence 2–14.[3]

Notwithstanding the above-mentioned implementation difficulties, Gazdar, on the assumption that all the sources of presupposition can be written as functions, defines f_p, the pre-supposition function which yields all the potential presuppositions of a sentence:[4]

[2](See Stalnaker [52] for another review of Gazdar's work).

[3]He admits in a footnote that "this is insufficient, since most factives also presuppose that the subject of the matrix sentence knows the complement to be true...". With this proviso and with the change from a knowledge (**K**) operator to a belief-predicated expression of the form employed later in this thesis, I am in basic agreement with this approach.

[4]The equations in the definitions employ Gazdar's original notation and terminology.

Definition 2 (Gazdar: pre-suppositions) *for any sentence* ϕ,

$$f_p(\phi) = \cup_{f \in F} f(\phi)$$

Gazdar also accounts for other inference classes, notably various implicatures, but does little to specify the functions that would generate them. He also admits that his theory is liable to the charge of *ad-hoc*ness, as the order in which the rules are applied is not argued for.

Briefly, pre–suppositions and im–plicatures become presuppositions and implicatures only if they survive the mechanics of Gazdar's 'satisfiable incrementation' system. In particular, this mechanism prevents the passing through of pre-suppositions which 1) should not project from the clauses of complex sentences into the set of presuppositions of the matrix sentence, 2) are inconsistent with the existing context, and 3) are also implicated or entailed by the sentence.

Example 2–6:

Sentence 2–15: If John sees me then he will tell Margaret.

Sentence 2–16: I don't know that John will see me. ■

In example 2–6, Gazdar gives sentence 2–16 as an example of a clausal quantity implicature of sentence 2–15. So in particular, the set of clausal quantity im-plicatures for simple disjunctions or conditionals is given by definition 3, where **P** is read as *for all the speaker knows it is possible that.*

Definition 3 (Gazdar: Clausal Quantity Im–plicatures)

$$f_c(\phi \frown or \frown \psi) = f_c(if \frown \phi \frown then \frown \psi) = \{\mathbf{P}\phi, \mathbf{P}\psi, \mathbf{P}\neg\phi, \mathbf{P}\neg\psi\}$$

Mercer sets out to formalize certain presuppositional phenomena within the framework of a default logic[39].He recognizes the crucial importance of the defeasibility of the presupposition relation, and takes this as persuasive evidence for modelling it within a default logic. He identifies three distinct sources of presupposition defeat: contextual, conversational, and where propositions which are presupposed by a sentence are also entailed by it. These desiderata, along with the behavior of the relation under negation, lead to Mercer's proof-theoretic definition of presupposition:

Definition 4 (Mercer: Presupposition) *A sentence* α *is a presupposition of an utterance* **u**, *represented by the default theory* Δ_u *iff*

- $\Delta_u \models_\Delta \alpha$ *and*
- $\alpha \in Th(CONSEQUENTS\{D\})$, *but*
- $\Delta_u \not\models \alpha$ *and*
- $\Delta_u \not\models_\Delta \neg\alpha$ [5]

[5]Mercer notes about this definition that: "...the only defaults in Δ_u are the presupposition generating defaults. In reality the default theory would contain many other kinds of defaults. The definition would have to be changed so that the proof of α requires the invocation of a presupposition-generating default ...As well, ...all proofs must require the use of the statement representing the semantic representation of the uttered sentence." Similar considerations motivate aspects of the implementation presented in chapter 3 of this thesis.

The technique is far more principled than its precursers. It no longer suffers from the form of *ad-hoc*ness attributable to Gazdar's theory, but there are other *ad-hoc* steps in the derivations of presuppositions from default theories representing complex sentences; this problem leaves large question marks for anyone interested in a working implementation of the method, but Mercer's remains the most principled approach, and my work on presupposition follows closely on his. Refer to § 4.7 for further comparison of Mercer's approach with my own.

Horton has recently presented another theory of presupposition, with an emphasis upon modelling presuppositions as beliefs of agents [26]. In particular, she points out that not only do the beliefs of the speaker and listener have to be accounted for, but that the beliefs of other agents need sometimes be included to provide an intuitively satisfying account of the presuppositions of some complex sentences.

Horton also gives much consideration to the defeasibility of the presuppositional relation, carefully distinguishing between presuppositions which are *blocked* by semantically internal negation, and those which must be *retracted* due to inconsistency with antecedently or subsequently established context. Horton's potential presuppositions resemble Gazdar's *pre-suppositions,* although she is careful to point out that when a sentence potentially presupposes a proposition, that sentence tends to imply that proposition. She is therefore attaching more than mere 'technical' significance to potential presuppositions. Horton agrees with Gazdar that the 'survival of candidate presuppositions depends on consistency'[26]. The reader will note in subsequent sections of this thesis, that my approach is entirely consistent with Horton's belief-centered view of presupposition, though my theory has different goals than hers. I am interested in developing a theory of communication, to which beliefs of the interlocutors are crucial; Horton also recognizes the importance of agent beliefs, and her theory of presupposition is also couched in terms of these beliefs.

Burton–Roberts One might think or at least hope that after so long a history, some sort of consensus would have been reached on what presuppositional phenomena are, and upon how they behave. Remarkably, not even the semantic–pragmatic division has been surmounted, as evidenced by the recent publication by Burton-Roberts [8], which claims nonetheless to prove once and for all that presupposition is semantic in nature. Burton-Roberts' approach is well motivated. I concur with him on both his dissatisfaction with semantico-logical definitions of presupposition, and with his observation that "projection problems are thrown up by definitions. Without a definition, there can be no problem." He then frames the project in terms of what he sees as three misguided assumptions pervasively manifested in previous work.

In sum, the motivation for his approach—like Strawson's—is the desire to explicate a purported intuition about truth-valuelessness in sentences which admit of presupposition failure. His approach relies on a bivalent logic with gaps, of which he concludes he has "no means of conclusively demonstrating that the distinction between trivalence and gapped bivalence consists in what I say it consists in." He is able to handle a large range of examples that have been classically problematic for standard semantic theories of presupposition, but the approach is essentially Strawsonian.

His technique appears to work for those examples which Mercer and others have identified as problematic for semantic theories, but as noted, Burton-Roberts does give up bivalence, and his motivations are different from mine. In shopping around for a definition of presupposition, motivations are relevant; he has very little to say about cancellation of presupposition by contradictory context. Some 'pragmatic', context-sensitive mechanism is required above and beyond even a successful semantic account of (truth-functional) projection.

2.1.2 Presuppositional Environments

Burton–Roberts [8, p249] credits Rob van der Sandt (in conversation) with the observation that "*every* theory is, in the final analysis, going to have to list the presupposition-inducing elements anyway." Gazdar has provided only a hint of how this might be accomplished via his pre-supposition generating function f_p, reproduced herein with definition 2. Mercer [39, p34] lists a range of environments which carry presuppositions, and formulates some of these within a default logic. It is implicit in his work that although he has presented only *some* of these environments, it is possible in principle to list them all. Karttunen, has listed thirty such environments.

Horton [26, p71] also lists a selection of presupposition-carrying environments, which she calls *triggers.*

In short, the consensus appears to be that there is a finite number of presuppositional environments, and that they can all –in principle– be enumerated. No one to my knowledge has made a claim as to the number involved. The theory presented in this thesis makes these assumptions as well.

2.1.3 Summary

Much of the work I have reviewed in this section seeks to provide an account of the truth-conditionality of sentences which exhibit presupposition failure. Thus, while Burton-Roberts' theory may succeed on this count, its usefulness to my project without some explanation of context-incrementation is limited. My project is the derivation of agents' beliefs from utterances, which must take into account much of the doxastic environment of the agents which are involved.

Perhaps the greatest difficulty for proponents of a semantic approach to presupposition is the so-called *projection problem.* This is the study of how the presuppositions of the clauses of a complex sentence 'project' over the sentence and into the context. Burton–Roberts [8] has argued that projection has been a problem for semantic approaches only because previous definitions of presupposition adopted by semanticists have been incorrect, and presents his own version. He also argues that the perceived ambiguity of natural language negation is likewise a by-product of a misconceived semantic definition.

Pragmatic approaches avoid these issues largely by sidestepping them, and derive their (considerable) explanatory power from high-level theories of communications, although this is not always explicated. The result is that there is still no well-principled account of presupposition which can be derived strictly from a theory of communication. I go on to propose an axiomatization of such a theory, along with the necessary rules of inference to derive not only presupposition, but other classes of pragmatic inference as well.

Dimension	*Research*					
	Russell	Strawson	Gazdar	Mercer	Horton	Csinger
Logic	Classical	3-valued	ad-hoc	default	modal	**Theorist**
Defeasibility		√	√	√	√?	√
Implementation						√
Belief-predicated			√?		√	√
Context–sensitivity			√	√	√	√

Table 2.1: Summary of Previous Work in Presupposition

It may be quite simply that those who are most concerned with truth-functionality in natural language are forced into a semantic account of implicit as well as explicit phenomena, while those concerned most with the effects of context will tend toward pragmatic approaches as the most profitable tools.

While much of the previous work on presupposition centered on the attribution of truth values to sentences, I am more concerned in this thesis with the beliefs of agents in varying presuppositional environments. I have taken an 'opportunistic' approach to the use of presupposition within the implementation; wherever it seemed to me that additional beliefs could be derived from utterances, I implemented a presupposition schema to do my bidding. Thus, I make no claim that the range of presuppositional phenomena exploited by the system is complete; far from it. My approach does assume, following Karttunen and others [29], that "The basic presuppositions of a simple sentence presumably can be determined from the lexical items in the sentence and from its form and derivational history..." and that it is possible to "give a finite list of basic presuppositions for each simple sentence of English."

Table 2.1, a summary of the previous research in presupposition evaluated in this section, provides an at-a-glance statement of particular researchers' attention to the dimensions I have identified. A question mark in any box indicates my feeling that the dimension, though mentioned in the work, is tangential to the thrust of the research.

2.2 Theories of Communication

Several approaches have been taken to theorizing about, or modelling communication. All are subject to the validity of numerous assumptions, and none of them has been completely adequate. These are some of the issues which any theory of communication must address, and which are briefly dealt with in the following sections:

- The meanings of utterances
- The mechanism(s) which support(s) the derivation of meanings of utterances
- The *purpose* of communication
- The organization of knowledge

Utterance Meaning A well-established view [9] is that the meaning of a natural language utterance consists of the logical form of the utterance itself, along with all of the inferences that can be made from this logical form and any relevant, available knowledge.

This view remains plausible, but too vague to be more than a guideline. It makes no claim as to the nature of the logical form, the inference method, or categories of knowledge required. Various formalizations exist, which are more committed along one or more of these dimensions. In general, some distinction is made between the propositional content of an utterance, and the meaning of an utterance; the former is a subset of the latter. Diversity of terminology is a factor here as well; Herzberger [25] has referred to propositional content as *assertive* content. **Horton** has suggested that the *communicative content* of an utterance includes its entailments, conversational implicatures, conventional implicatures, and its presuppositions [26, p1]. **Gazdar** has distinguished between *literal* and *conveyed* meaning, and provided rules and conditions for deriving the latter from the former. **Mercer** writes that his model of communication rests upon two assumptions; the first of these has to do with cooperative principles, while the second is concerned with sentence meaning. He suggests that "...the *meaning* of an asserted declarative sentence is approximately equivalent to *update your knowledge base with the logical form of the sentence just uttered.*[39]"

Bach and Harnish [5, p150] discuss issues of sincerity versus 'literalness,' intending versus operative meaning versus Grice's notion of speaker meaning.

Deriving Utterance Meaning While it is generally agreed that the meaning of a sentence is more than just its propositional contents, as noted above, there is not much consensus upon what this meaning actually consists of, much less any agreement about how to derive it.

If, as Marr says, "...phrasing of information must be an artwork of suggestiveness and insight", then the retrieval of the information must be via a process that is equally sophisticated. Gazdar [20, p133], for instance, suggests various levels beyond the *literal* meaning of an utterance; within the system of *satisfiable incrementation* which he defines with some precision, the relevant quantities are *conveyed meaning,* and *conversational contribution,* defined as follows:

> The *conversational contribution* of an utterance ...is that proposition ...which consists of all worlds *except* those that have both the following properties:
>
> - they were included by *all* the propositions in the context of [the] utterance and
> - they are each excluded by at least one proposition (not necessarily the same one in each case) in the context that results from the utterance.

Mercer [39] defined sentence meaning as approximately equivalent to updating the hearer's database (above). He adds that implied in this is a "commitment to the principle that the inferences are generated by a well-founded proof theory working in conjunction with knowledge represented as statements in a logical language." Mercer deals only with asserted declarative sentences, and only with the generation of their presuppositions. The project then becomes one of defining the presupposition relation, which he goes on to do within a default logic formalism.

The Purpose of Communication Linguists, philosophers, and -recently- computer scientists have grappled with the nature of communication. How is it accomplished? What makes it possible? Certain assumptions have been at the heart of all theories

thus far, often referred to as *principles of cooperation.* These amount to no less than normative guidelines for communicative acts and processes. Mercer's [39, p7] first self-professed assumption regarding his model of cooperation is that "...the rules given in Grice's theory of cooperative communication govern the communication act." Neither is Horton [26, p30] free from these assumptions. Her introduction reads:

> Before proceeding, we now pause to discuss the assumptions that we make. In order to simplify the problem, we will follow Grice in assuming that conversation is cooperative. Specifically, we will assume the following:
>
> - Sincerity Assumption: The speaker will only say what he believes to be true. In other words, the speaker will not deliberately try to deceive the listener.
> - Straightforwardness Assumption: The speaker will not use sarcasm (a flouting of the maxim of Quality).

Although she goes on to suggest how her assumptions might be relaxed in order for her theory to model deceit and sarcasm [p96], hers is not a general theory of misleading.

Organization of knowledge Although the enthusiasm of early researchers in knowledge representation is reflected in the terminology that still pervades the area, I do not see any gain in clarity via the use of such terms as 'knowledge base,' 'rationality module,' 'PLANNER,' 'Conniver,' etc. [42]. While such names are intentionally idiomatic, and highly suggestive of the roles they play in the (toy) implementations of their creators, they obscure the huge gulf between what they are and the psychological analogs they are designed to emulate. I suggest then, that the following categories of *information* are salient.

- Situational Information: Mercer distinguishes *situational* from *background* knowledge [39, p11]. In particular, he writes that utterances are an important source of situational information, and thereafter restricts himself to only this form of situational information.He suggests that other sources might include information from previous parts of the discourse, the physical situation of the interlocutors, and their relative social statuses. As far as the current work is concerned, I too am interested only in the utterance itself as a source of situational information.

- Background Information: Background information is everything non-situational. Aside from this obvious description of what it is not, various implementations have categorized it in different ways. I identify the following categories of information:

  - Linguistic Information: generally includes the following sources of information: phonology, syntax, semantics, pragmatics. The concerns of this thesis do not touch upon issues of phonology, and the domain can be safely ignored. Syntactic information is usually represented in the form of a *grammar,* which is used to build a structural representation of the utterance. Various formalisms exist, and their output is treated by some semantic process to yield a *logical form,* corresponding to what I have labelled the propositional content of the utterance. Semantic information is sometimes classified as including such facts

as, for example, that factive verbs like *regret,* and *surprise* entail their complements [39, p12]. Although not crucial to my arguments, I go along with such categorizations. Pragmatic information encompasses a vast (and nebulous) area that includes information about the conversational usage of language in different situations.

Of particular interest under the heading of (pragmatic) linguistic information is the aforementioned Cooperative Principle.

- World Information: World, 'real-world,' or 'encyclopaedic' information includes facts about the world as it is. I take to be under this heading such 'knowledge' as the binary quality of human sexuality (*i.e.,* that humans are generally male or female, and that these states are generally mutually exclusive), the 'knowledge' of the capital cities of the world, and so on. Information in this category can be of a default nature as well; some people *are* hermaphroditic.
- Contextual Information: Human communication—and indeed the human concept of understanding—is grounded in context. Whether an utterance is successful is measured against the change it produces in the beliefs that the hearer has of the world; the speaker also has beliefs, and some of them are about the hearer. These beliefs are always subject to revision, and thus hint again at defeasibility from another direction. *Context* is conventionally regarded as the set of beliefs that are shared by the interlocutors. It has been called *shared knowledge, mutual knowledge, common ground,* etc.

2.2.1 Principles of Cooperation (Grice)

All of the previous systems have employed some version of the Cooperative Principle developed and summarized by Grice [22], and repeated here as Figure 2.1. In its simplest form, the principle accepts that the semantics of the language is *a priori,* and that utterance meaning depends upon this semantics augmented with inferences sanctioned by rules describing conversational use of utterances. These rules comprise the (Gricean) *cooperative principle:*

> Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.

Whether this principle owes its longevity to only its vagueness, or to some other quality, the fact remains that in some guise or another, it underlies every model of communication that I have surveyed [22, 5, 20, 23, 38, 39, 27, 8]. The Cooperative Principle captures what I have called *The Assumption of Minimal Perversity.* This is the element of reasoning which eludes monotonic logics, and what all non-monotonic systems attempt to capture.[6] In this study, minimal perversity manifests itself in that *given no indications to the contrary* the hearer assumes that the utterance adheres to the reasonable guidelines of the

[6]I have noted the following definition of the Minimal Perversity Assumption: the assumption that of the (possibly, or even likely) infinite number of clauses which might affect the reasoning process, only those whose truth value is known must be considered. This is analogous to the well-known *Closed World Assumption,* and to various circumscriptive devices, all suggestive of non-monotonicity.

1. Quantity
   - Make your contribution as informative as is required.
   - Do not make your contribution more informative than is required.
2. Quality: Try to make your contribution one that is true.
3. Relation: Be relevant.
4. Manner: Be perspicuous.

Figure 2.1: Grice's Maxims of Conversation

Cooperative Principle, embodied by the maxims of Figure 2.1. Typically, these maxims are applied from a strictly extra-logical perspective, although various re-formulations are in use.

2.3 User Modelling

User modelling is the investigation of how assumptions about a user's background knowledge (as well as the user's plans and goals in consulting the system) can be automatically created, represented, and exploited by the system in the course of interaction with the user [31, p3]. User modelling is a special case of *agent modelling*; this thesis presents an approach to agent modelling in a natural language environment which is based on default reasoning. Others have considered goals and plans of the user [12, 2]; I restrict myself here to the user's *beliefs*. Although this may be a good point at which to launch into what beliefs actually *are*, I will pay these philosophical dues in § 2.4.1.

User modelling takes many forms. One approach has been to let the user do the modelling; some applications permit the user to modify variables in the environment which reflect user-proficiency parameters indirectly [14], while others query the user directly about preferences and capabilities. If user-modelling has come a long way since a user-determined 'help-level' first appeared on the menu of a popular word processing program, it has much farther to go before achieving the kind of flexibility we expect from our human interlocutors.

User modelling is what interactive systems will need to do to be responsive to the needs of the user. Tutorial systems need to guage the user's competence in the subject of study [28, p184]. The earlier discussion about the meaning of an utterance is relevant here (§ 2.2). Presuppositions and other non-propositional utterance-related phenomena have recently resurfaced in the 'computational' literature because of their connection with database systems in general, and question-answering in particular. There has been some debate over what constitutes a cooperative response from a system when the question put to it suffers from presupposition failure [40, 28]. Kass and Finin [31] have enumerated a number of dimensions along which user-models can be categorized. Of particular interest are the dimensions they call *Representation of beliefs* and *Acquisition of beliefs.*

Representation of Beliefs Of interest under this heading is the continuum between what Kass and Finin call *implicit* and *explicit* representation. Systems which model attitudes implicitly are of little interest to my project. These include any implemented program in which the programmer has made any assumptions about the prospective users of his system, while writing the code. (Kass and Finin cite a generic FORTRAN compiler as an example.) Explicitly encoded attitude representation has some characteristics, some of which reflect on the project at hand. Part of what I have been referring to as *Rationality,* Kass and Finin call *explicit representation,* and go on to say:

> The knowledge in the agent model is encoded in a representation language that is sufficiently expressive. Such a representation language will typically provide a set of inferential services, allowing some of the knowledge of an agent to be implicit, but automatically inferred when needed.[7]

Acquisition of Beliefs The method by which beliefs are acquired is relevant to the effectiveness of the user-modeller [31]. Acquisition has been described with the already overused terms *implicit* and *explicit.* Explicit acquisition takes place when the user makes explicit statements about what she does and does not believe. Implicit acquisition is more difficult, and involves deduction on the part of the user-modeller. One approach is the technique advanced in this thesis, wherein the user-modeller monitors the communications channel between user and application, and derives tacit and implicit user-beliefs from the propositional contents of user-utterances.

Belief acquisition in the domain of user-modelling has been further categorized [ibid.] as *recognition oriented* or *constructive.* Recognition oriented approaches are more limited in their scope, but more straight-forward to implement. This kind of system relies on a stored set of belief *stereotypes*, which can be triggered by the form of a user-utterance [ibid]. The approach to belief acquisition advanced in this thesis is thus implicit, and constructive.

2.4 Belief and Rationality

Beliefs and rationality are deeply intertwined, both within and without the computational paradigm. In the following subsections, I explore the relationship between belief and rationality, and consider various definitions of rationality. The aim of this investigation is to suggest directions that might lead to plausible formulations of belief introspection (*i.e.*, rules to derive beliefs from existing beliefs).

In this thesis, I take the working view that

1. An intelligent agent can be described in terms of its beliefs[8]
2. Rationality is in some sense a *well-formedness criteria* for the beliefs of an intelligent agent

[7] *cf.* Levesque's distinction between *implicit* and *explicit* belief [35].

[8] Though I have not completely abandoned this original, naive hope that goals and desires might be expressed as complexes of beliefs, it certainly appears to me now that it is *easier* to treat them as primitive.

The first point claims –without trying to explicate the nature of a belief, (*i.e.* without making any ontological claims in respect of beliefs)– that an intelligent agent can be described at some level by its beliefs; for instance, one agent can be distinguished from another by a difference in their respective beliefs. In general, I will refer only to a partial description of this sort. Thus, when I speak in terms of an agent's attitudes, goals, or desires in this thesis, I do so loosely, with the underlying assumption that these aspects of the agent's mental state can ultimately be reduced to some expression in terms of its beliefs.

The last point is the one that ties beliefs to rationality. Agents that exhibit an identifiable set of normative characteristics are predictable to the extent of their 'normativeness'. In general, inter-agent communications relies on this normative component, and deviance results in various pathologies (*e.g.* pluralistic ignorance and false consensus). In particular, the theory described in this thesis provides predictive power for agents which are normative with respect to a particular definition of rationality, to be described. Hearer-agents with this normativity can make inferences about the beliefs of speaker-agents who make utterances, and normative speaker-agents can derive the forms of their utterances from their beliefs. This is the view of communication taken in this thesis.

Previous views of this well-formedness criterion have amounted to anything from classical logical consistency to *ad-hoc* procedural specifications. I try to leave the definition as loose as possible, to be filled in with further results as necessary, but I do propose herein that a default reasoning framework offers immediate results towards resolving some well-known problems such as logical omniscience.

In this section, I pay my philosophical respects to others who have considered the relationship between rationality and belief.

2.4.1 Beliefs

There is much to say about beliefs, as the ample body of philosophical literature demonstrates, but there is very little that is not still under investigation. The researcher who wishes to represent beliefs in a computational environment has little more than his own intuition to go on. My own intuition urges me to remain as ontologically uncomitted as possible, and while I have surveyed a wide range of models of belief, I will stay with what I consider to be the most minimal.

Beliefs ***versus*** **Knowledge.** I will speak only in terms of the *beliefs* of the agent being modelled. Other approaches in which the subjectivity of truth has been recognized are current, *viz.* [51]:

> ...start with only a definition of knowledge, any definition that you find acceptable, and define belief as a defeasible version of it.

'Beliefs' are common-sense entities, and become the objects manipulated by a default logic in virtue of just this quality. Logical definitions of mutual belief have been offered by many researchers. The subject is of some importance to this project because the Cooperative Principle in operation demands mutual recognition, and hence representation. Suffice it to say that the interlocutors postulated for my theory *mutually believe*[9] the elements of the

[9] All the previously discussed reservations about 'belief' apply here.

Cooperative Principle. If they do not, a state of pluralistic ignorance or of false consensus might arise, in which the maxims of cooperation would be defeated. Agents necessarily maintain models of their peer-agents; part of this model usually includes beliefs to the effect that, among other things, their peer-agents believe the maxims of the Cooperative Principle. (The unfoundedness of such a belief is characteristic of false consensus).

The Epistemic Status of Belief: Hadley [24, p4] surveys the epistemic status of beliefs in artificial intelligence research:

> We may summarize the stance towards belief currently adopted by many (though not all) AI researchers as follows: "Agent X believes sentence S if and only if S is explicitly present in X's belief base, or S is derivable, by means of a *tractable* epistemic logic, from a set of epistemic formulae corresponding to a subset of X's explicit belief base." ...Nevertheless, agents often have inconsistent beliefs, and do not automatically 'commit to' the conclusions they derived.[10]

Compare Konolige [33]:

> ...a belief subsystem is the computational structure within an artificial agent responsible for representing his beliefs about the world ...A belief subsystem consists of a finite list of facts the agent believes to be true of the world (the base set) together with some computational apparatus for inferring consequences of these facts ...the belief set of an agent is the set of all queries that can be derived.

The effort that has been devoted by philosophers of mind to the question of belief should not go unnoticed. In the absence of a computational workbench, some inquirers have constructed models which bear a striking resemblance to the architecture of the system presented in this thesis, and are useful exemplars of the kind of behavior I would like an artificial agent to exhibit. The work of cognitive scientists serves then, as a high-level *requirements analysis* for researchers whose aim is to implement cognitive models.

Stephen Stich's *Content Theory of Belief* [53] is one such model. In it, he proposes an *Inference mechanism* whose resemblance to the Rationality or Introspection module of this implementation is obvious. He considers beliefs to be some form of mental sentence tokens, whose meanings are imbedded in their causal interactions with other sentence tokens and with the environment. To connect the agent with his environment, he adds perception and action-control units. To deal with the causes of actions, he includes a Practical reasoning mechanism, capable of generating desires from beliefs and desires.

There is more than passing interest in this model, for it shows us how far a computational system must go before it can be considered to have beliefs in the same sort of way that we do. In addition to representing beliefs and providing a mechanism for introspection, it must take account of desires, and provide both a means to generate desires, and a way to interface with the environment. The notion of artificial agent in this thesis falls far short of these requirements.

Stich's model nicely points the way to future work. While we have developed a hearer-based computational theory of communication, and an implementation that has application as a user-modeller, we have ignored goals and desires, and therefore a realistic account of a speaker model is still out of reach.

[10] Hadley uses but does not explain in his paper what he means by *rational*.

2.4.2 Rationality

I take rationality to be the mechanism whereby an agent reasons about its world, generating new attitudes in concert with its previous attitudes and with incoming data from its environment (the context), and discarding any attitudes which it finds untenable in its system for rationality (*i.e.,* it must discard those attitudes which are *ir*rational). This circular definition leaves open many issues, and particularly the question of the mechanism itself. But controversy begins with the smallest move towards a more detailed account.[11]

Most work to date has tended to an often unexplored assumption that rationality is and must be at bottom *logical.* This belief has made its way into even the lay world where–accompanied by admonishments from Mr. Spock on the bridge of the Starship Enterprise–irrationality and illogic are confused. To avoid crossing unnecessary ontological territory, ground that is likely someday to be lost to a better-founded theoretical assault, I want from the outset to clearly separate what is the domain of rationality, and what part logic is to play in it.

A goal or a belief can be rational only with respect to an agent and his inference mechanism. For instance, it may appear *prima facie* rational for an individual to plan to have children, if the axiomatization of that agent's beliefs includes only his built-in (innate) instincts. But if the additional constraints of global population density are added, such a plan is questionable in its rationality.

Still, the inference mechanism remains unspecified, and the haste with which researchers join to define it in terms of first-order-predicate-logic is forgivable only in view of the shortage of viable rivals for the job. For the sake of the current computational implementation, I too follow the trend of exploring different logical axiomatizations of rationality. My approach is detailed in § 4.5.

Within the 'computational' school, established rationality constraints are usually some variation upon a demand for logical consistency, which is in my view an unrealistic attitude. I would not want to go so far as to suggest that (logical) consistency is a prerequisite for rationality, much less that (logical) closure be a criterion of rationality.

Previous strategies have all suffered from what has been called the problem of *logical omniscience,* wherein an agent who believes α is held to believe all of the (logical) consequences of α. This requirement imposes at least the following conditions [35]:

- Every valid sentence must be believed
- If two sentences are logically equivalent, then one must be believed if the other is (regardless of its complexity)
- If a sentence and its negation are both believed, then so must *every* sentence

These conditions are undesireable as partial definitions of rationality. I want something less limiting, and turn first to default reasoning for both a defeasible version of closure and a language capable of expressing inconsistency. I will also explore the (partial) implementation of implicit versus explicit belief along the lines of Levesque [35]. See § 4.5 for my implementation of the constraints on rationality.

[11] See [10] for a thorough discussion of these and related issues.

2.4.3 Previous Work in Belief Modelling

In this section I survey previous work in the modelling of beliefs, with particular attention to the aspects discussed above.

Allen [2, 1] has advanced a theory of speech acts, along with an implementation that makes use of rules composed of *preconditions, bodies,* and *effects.* The preconditions serve to embed the rules within consistent contexts, and enforce the Gricean maxims along the way. Allen says as much in his description of the INFORM speech-act [1, p443]:

> As expected, there is a precondition that the speaker believes the proposition that is asserted, and the effect is that the hearer believes the proposition.

The definition is given [2] in terms of shared beliefs of the speaker and hearer, and with the addition of the preconditions.

He defines other speech-acts (*e.g.,* REQUEST), whose effects or preconditions involve the *intentions* of the agents involved, *viz.,* their plans and goals. Allen notes that an accurate account of beliefs and intentions would need to be time-indexed. Other operators are introduced to represent further intentions.

Allen recognizes the limitations of the belief-logic he employs, which he says is to be interpreted more or less along the lines of Hintikka.

Only the propositional contents of formulae are considered in utterance meaning.

The point most salient to my project is Allen's recognition of the importance of context-sensitivity, which he implements via the preconditions of his operators.

Cohen and Perrault [12] also describe a system that makes use of rules consisting of preconditions and effects. They implement the INFORM and REQUEST speech-acts, and they interpret belief as a modal operator constrained with an axiomatization, which they recognize as an 'idealization' that is clearly 'too strong to be a faithful model of human beliefs'. **Cohen and Levesque** provide a set of context-sensitive axioms [11] to capture the consequences of utterances. Their approach makes use of a form of the *closed world assumption,* in that the preconditions for some of these rules involve statements about what an agent does *not* believe. Only the propositional contents of formulae are considered. Other operators are introduced to represent the intentions of agents.

Perrault addresses the application of default logic to a speech act theory [41], and in so doing brings many issues to light.

He realizes and argues strongly for the context-sensitivity of the rules that capture the consequences of utterances, and this is a large part of his appeal to non-monotonic logic.

Perrault deals exclusively with the propositional contents of declarative utterances. He distinguishes between knowledge and belief.

Perrault's approach to rationality makes no appeal to default logic. He states that "The beliefs of one agent at one time are taken to be consistent, distributive over conjunctions, closed under logical consequence and positive introspection. Beliefs need not be true." The strengths of his axioms prevents the revision of belief. Also, he indicates that all agents are assumed to believe that all axioms hold. This is not a default rule:

Definition 5 (Perrault: Axiom Closure) *For every agent x, time t and axiom A above, $B_{x,t}A$*[12] *is an axiom.*

The default rules employed by Perrault both concern the incrementation of belief sets.

Definition 6 (Perrault: Belief Transfer Rule) $B_{x,t}B_{y,t}p \Rightarrow B_{x,t}p$

Definition 7 (Perrault: Declarative Rule) $DO_{x,t}\mathbf{p} \Rightarrow B_{x,t}p$

The *Declarative Rule* is similar to the sincerity condition which has been referred to throughout this thesis, and which will also be implemented in § 4.2 as a default rule. $DO_{x,t}\mathbf{p}$ is to be interpreted as *the action of agent x at time t of uttering a declarative sentence with propositional content* p.

He adds the following meta-rule, to implement closure of default rules:

Definition 8 (Perrault: Default Rule Closure) *For all agents x and times t, if $p \Rightarrow q$ is a default rule, so is $B_{x,t}p \Rightarrow B_{x,t}q$*

The implementation *within a logic*[13] of rules such as these is always a problem. (See § 4.5.)

There is no sense in which one extension of a default theory has precedence over another. I will discuss this issue in § 2.5.2.

Perrault is able to show in this formulation the difference between theories representing sincere and insincere utterances by a speaker, but the persistence axiom (as noted above) prevents the retraction of previous beliefs. He briefly considers different belief strategies that might be described in default logic, pursuing the possibility of making some of the axioms into default rules. In particular, he mentions the persistence axiom and discusses its conversion into the persistence default rule. If both this and the memory axiom were converted to default rules of inference, multiple (mutually inconsistent) extensions would result, representing both the case in which the hearer's beliefs persist, and the case in which they do not. As he points out, "The theory would then give no precedence to either." Perrault does not pursue the subject any further, though it seems to me that this is the single most important unexplored thread. See § 2.5.2 for further exploration of the problem of choosing between multiple extensions.[14]

Konolige [33] and Batali [6][15] both explore the ability of agents to reason about their own representations, a process that they call *introspection*. Konolige advances the view that 'a belief subsystem is the computational subsystem within an artificial agent responsible for representing his beliefs about the world.' Konolige [32] argues that this belief subsystem is 'conceptually separate' from the rest of an agent's cognitive mechanisms. He also distinguishes between the finite list of facts which the agent believes *a priori* to be

[12] $B_{x,t}P$ is read as *agent x believes that P at time t.*

[13] It is easy enough to add *meta*-logical control to implement rules like this, but these are variously *ad-hoc*.

[14] Perrault goes on to consider *intentions* within the limits of his formulation, but these issues fall outside the scope of this thesis.

[15] Batali's work is a survey of several computational models of introspection, and includes an extended argument for continuing such research. Most of his observations are covered in this thesis in some form or another.

true of the world, and the set which the agent can derive via its computational inference apparatus. He calls the finite set of beliefs the *base set,* and the inferrable superset the *belief set*; these notions correspond closely with Levesque's *explicit* and *implicit* belief, respectively.

Konolige, with Appelt [3], also advocates the use of default logic, with emphasis upon attitude revision. He employs what he calls a *hierarchic autoepistemic logic,* characterized by a collection of subtheories linked in a hierarchy, rather than by a single default theory. I will have more to say about this approach in § 2.5.2.

2.5 Non-monotonic Systems

Default logics were formulated to overcome some of the well-known problems of classical, monotonic logics.

> **Definition 9 (Monotonicity)** *A system is* monotonic *if and only if it has the following property: whenever it infers a conclusion C from a set of assumptions S, it will also infer C from any larger set of assumptions containing S.*

One of the best known non-monotonic formalisms is due to Reiter [49].

2.5.1 Theorist

The **Theorist** formulation for default reasoning lends itself particularly well to implementation in a logic programming environment [45]. The **Theorist** implementation I used embodies a non-clausal first-order theorem-prover, and a mechanism for defeasible rules of inference, making it a likely candidate for implementing both the principles of cooperative communication, and the rules for presuppositional inference.

In **Theorist** the user provides two sets of first order formulae

$\mathcal{F}$ is a set of closed formulae called the *facts*. These are intended to be true in the domain being modelled, and as such are assumed to be consistent.

Δ is a set of formulae which act as *possible hypotheses*, any consistent ground instance of which can be used as a premise in a logical argument.

> **Definition 10 (Scenario)** *a* **scenario** *of $(\mathcal{F}, \Delta)$ is a set $D \cup \mathcal{F}$ where D is a set of ground instances of elements of Δ such that $D \cup \mathcal{F}$ is consistent.*

> **Definition 11 (Explanation)** *If g is a closed formula then an* **explanation** *of g from $(\mathcal{F}, \Delta)$ is a scenario of $(\mathcal{F}, \Delta)$ which implies g.*

> **Definition 12 (Extension)** *An* **extension** *is the set of logical consequences of a maximal (with respect to set inclusion) scenario.*[16]

> **Definition 13 (Prediction)** *g is* predicted *if and only if g is in all extensions.*

That is, g is explainable from $(\mathcal{F}, \Delta)$ if there is a set D of ground instances of elements of Δ such that

[16]This corresponds to Reiter's definition of extension in terms of fixed points[49],[46].

$\mathcal{F} \cup D \models g$ and
$\mathcal{F} \cup D$ is consistent

in which case $\mathcal{F} \cup D$ is an explanation of g. Such a g will be referred to in this thesis as the *explanandum*[17] of a logical argument.

I will make extensive use of both prediction and explanation as described above, in the discussions to follow.

Theorist is an attempt to be a minimalist system. It is an attempt to see how far a very simple hypothetical reasoning framework can be pushed. It will also be of interest later in this thesis because exactly the same formal definition provides a definition for default reasoning, abductive reasoning, design, and recognition. These issues will arise in Chapter 3.

2.5.2 Theory Preference

The problem of multiple extensions arises in all default theories of any complexity. There is great representational power in being able to place into separate extensions mutually inconsistent formulae corresponding to distinct alternatives. This power has, however, gone unused because of the problems associated with choosing between the extensions.

The implementation presented in this thesis also suffers from the multiple extension problem, as will be detailed in § 4.7 Some comments by Perrault [41] highlight the difficulties:

> Ideally, one would like a theory in which it is possible for one agent's beliefs, say, to change depending on HOW STRONGLY[18] 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 will rely on something simpler. We assume what one might call a *persistence theory of belief:* that old beliefs persist over time, and that new beliefs are adopted as a result of observing external facts as long as they do not conflict with old ones.

Perrault has not gone into the reasons for his inability to provide 'such an account'; even the *ideal* theory he refers to does not address the discarding of beliefs in the light of new facts, and the problem of implementing looms large. This is no criticism of Perrault; his silence speaks eloquently for what needs to be done.

Time does not permit an exploration of the efforts thus far undertaken by researchers such as Poole [44], Brewka [7], Konolige [3], Geffner [21], and others, broadly characterized by a common goal of achieving a reasoning behavior in closer correspondence with intuition. Most approaches take recourse to some form of semantic, domain-dependent cues, thereby abandoning one of the stated goals of this thesis, that of ontological and logical minimality.

---

[17] The plural of this term is, of course, *explananda*!

[18] My emphasis

Chapter 3

Design Issues

Consistency is the hobgoblin of small minds.

—Ralph Waldo Emerson

I return now to my stated goal of deriving an agent's beliefs from his utterances. Having argued –as have many others– that the meaning of an utterance is more than its propositional, *explicit* contents, I go on to show how certain elements of the *implicit* and *tacit* contents play roles in the derivation of beliefs. The general model I will pursue is illustrated in Figure 3.1. The lines of the diagram indicate inference paths; the solid line between utterance and belief represents the familiar entailment relation of monotonic logic, while the other lines are intended to be suggestive of the defeasible implication of non-monotonic logic.[1] It is the purpose of this chapter to describe in some detail the inference processes which occur along these paths.

I embarked upon a default reasoning implementation not because I hold any measure of psychological reality for the formalism, but from the practically motivated desire to produce a system that would successfully ascribe a set of beliefs to an agent based upon his utterances. Some intentional idiom was needed, and default logic presented itself as the most accessible, the least *ad-hoc* and with the least ontological baggage.[2]

The strategy in all of what follows is to abstract away from the temporal linearity of discourse that would lead into truth-maintenance considerations, and to assume instead that the entire discourse is available for analysis. The problem is then one of achieving a *consistent explanation* of the discourse.[3]

I have already argued that it is advantageous for the system to make use of the entire bandwidth of the communications channel between user and application, and I

[1]Compare Figure 3.1 with Figure 1.1; the former can be interpreted as the elaboration of the latter, or the simpler Figure 1.1 can be seen as the limiting effect of a purely monotonic logic. The only inference path available to a purely classical analysis is the one solid line of Figure 3.1.

[2]Much of the remainder of this chapter appeared in [15].

[3]Perrault of SRI almost convinced me during a recent seminar that his default-logic formulation of speech-acts requires the use of time-predicated modal operators. A full description of discourse will require some reference to time, and perhaps even to modal operators, but it is my contention that the more limited project of determining the propositional contents of an utterance are well within the domain of minimalist (default) logic.

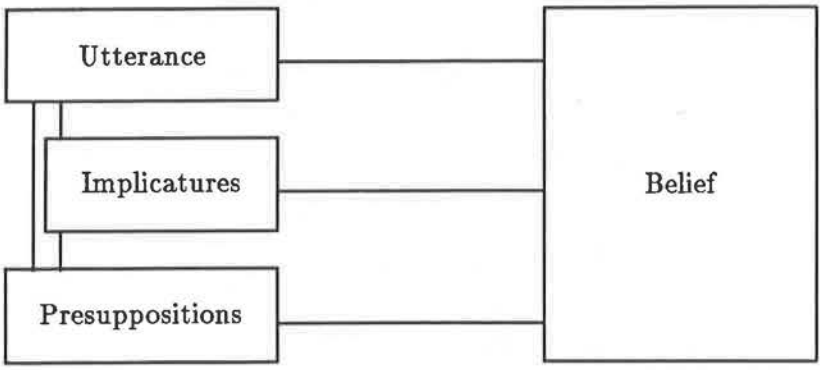

Figure 3.1: From Utterance to Belief via Communication

have suggested how this might be at least partially accomplished [17]. In particular, I have presented the inference classes of *presupposition* and *implicature* as part of the process by which to derive the *beliefs* of the speaker. This is the purpose of the theory of communication I have been advocating in this thesis.

3.1 Default–Programming Methodology

In general, there are not enough constraints in a domain to uniquely determine the approach that the reasoning system should take in formalising its characteristics [48]. The causality in the domain does not uniquely constrain its default-reasoning axiomatisation.

Different uses of **Theorist** can be characterized along two dimensions:

- Status of Explananda, and
- Status of Assumptions

The first considers whether the explanandum is known to be true or whether it is something that has to be determined. The second considers whether the system is free to choose any hypothesis that it wants or whether it must try to "guess" some hypothesis that "nature" has already chosen.

3.1.1 Status of Explananda

The first dimension is whether the explanandum is known or not. This choice corresponds to the following:

Abduction: The system regards the explanandum (the observation of the world or the design objective) to be true, and needs to find an explanation for it. The idea is to find assumptions that imply the goal. We consider all explanations of the goal as possible descriptions of the world.

Prediction: The system does not know if the explanandum is true, and the idea is to determine what can be predicted from the facts (the general knowledge and the observation or design objective).

		Explanandum	
		Known *Abduction*	Unknown *Prediction*
Who	**Design** User		
	Recognition Nature		

Table 3.1: Domain–Formulation

One interesting difference between abduction and prediction is in the relevance of counter-arguments. In predicting g, it matters whether or not $\neg g$ can be explained. In abduction, however, an explanation of $\neg g$ is irrelevant.

3.1.2 Status of Assumptions

Along the other dimension we can distinguish between the two tasks:

Design tasks are those in which the system can choose any hypotheses it wants. For example, a system can choose the components of the design in order to fulfil its design objective, or choose utterances to make in order to achieve a discourse goal.

A consistency check is used to rule out impossible designs. All other sets of components that fulfil the goal are possible, and the system can choose the "best design" to suit its goal. Design can be done abductively to try to hypothesize components in order to imply a design goal. Alternatively, design can be done predictively (*i.e.*, deductively) to derive a design from goals and any hypotheses we care to choose.

Recognition tasks are those in which the underlying reality is unknown, and all we can do is to guess at it based on the observations we make about it. This definition includes diagnosis, scene recognition and plan recognition. Recognition can also be performed abductively or predictively [46], [48]. In an abductive framework, we need to treat all of the explanations as possible descriptions of the world. In the predictive framework, an appealing strategy is to predict something only if it is explained from the observations even when an adversary chooses the hypotheses [47], which corresponds to membership in all extensions.[4]

Note that these frameworks are different ways to use the same formal system for different purposes. In order to use the system we have to choose one way to implement our domain.

3.2 The Communications Domain

Understanding is difficult even in the simplest of communications domains. Typically, a Hearer attempts to reconstruct a Speaker's (complex) mental state from a limited set of verbal and non-verbal cues, given only a general *a priori* understanding of the communications domain. The reasoning system with which we propose to implement inter-agent

[4]Which corresponds, propositionally at least, to circumscription [18].

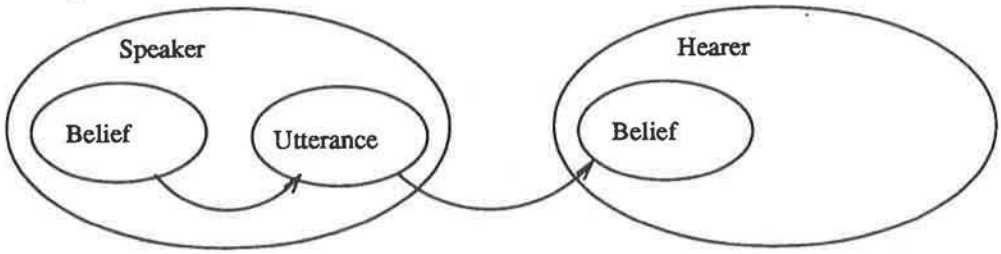

Figure 3.2: Causality Model for Interlocutor Pair

communications has only the utterances of some agents and a set of shared principles as input to the inference process.

A hearer-agent must resort to some form of theorizing about the speaker's mental states, based upon this sparse input. It is in such a frugal environment that the typical user-modeller must function, and for which the present theory is formulated: by 'listening in' to the utterances between interlocutors, the reasoner of the **UM** must reconstruct components of their mental states[5].

In the simplest **UM** system, the reasoner plays the role of hearer; based upon the utterances of the speaker, it attempts to reconstruct via some inference process, a subset of the mental state of the speaker [2]. As far as a speaker-agent is concerned, it is her intentions, goals and beliefs which compose the explanation of her utterance. The speaker's mental state can be regarded as *causing* her utterance, and this is likely to be the point of view of the speaker herself. The mental state of the speaker can be regarded as a representation of her *design objectives*; what she seeks is to *design* an utterance to fulfil these objectives.

As far as the hearer is concerned, it is the speaker's utterances which are the primary source of the hearer's beliefs about the speaker's mental state. Thus, the hearer seeks to *recognize* some components of the speaker's mental state from speaker-utterance. (Refer to figure 3.2.)

This exploration of the inherent direction of domain causality leaves open the direction of inference that the system is to select. This choice is essentially a question of default-logic programming methodology, since the way the domain is axiomatized will impose a particular inference strategy on both hearer and speaker agents.

3.3 Domain Formulation

We now turn to formulating the domain within the default reasoning framework. The problem of finding the right constraints on the domain breaks down into the problem of where to place the interlocutors of the speaker–hearer pair on the domain-formulation grid of table 3.1.1.

Elsewhere, we have discussed the kind of information needed to support interaction between rational agents, and have discussed specific points (*e.g.*, world knowledge, linguistic knowledge, and the extent to which these are *shared* by the interlocutors [17]).

[5]Kass and Finin [31] have referred to this approach to user-modelling as *implicit* with respect to acquisition, and *explicit* with respect to representation

		Explanandum (x)	
		Known *Abduction*	Unknown *Prediction*
Who	**Design** User	I. Speaker$_{11}$ $x = bel_s$	II. Speaker$_{12}$ $x = utt_s$
	Recognition Nature	II. Hearer$_{21}$ $x = utt_s$	I. *Hearer*$_{22}$ $x = bel_s$

Table 3.2: Communication Domain Formulation

Speaker		Hearer	
agent	uses	agent	uses
(1, 2)	prediction	(2, 2)	prediction
(1, 1)	abduction	(2, 1)	abduction
(1, 1)	abduction	(2, 2)	prediction
(1, 2)	prediction	(2, 1)	abduction

Table 3.3: Four Possible Implementations of the Domain

Philosophical issues aside, we suggest that in re-constructing a model of the speaker from her utterances, a hearer makes particular use of shared knowledge. To make this easier, the shared knowledge should be represented in a form that supports the inferences of both the Speaker (as utterance designer) and the Hearer (as belief recognizer). If we accept that there are principles of communication [22] which the Speaker adheres to in designing her utterance, it is reasonable that the Hearer make use of these principles as well during the recognition process. The central implementation question is then: *how should the principles of communication be represented?*

The answer to this question is hidden in an important characteristic of the interlocutor pair: *Speaker-Hearer Duality.*

3.3.1 Speaker–Hearer Duality

As we have presented the domain, there are essentially two kinds of information available to, and distributed between, the interlocutor pair: As a *designer* of utterances, the Speaker *knows* beliefs, while as a *recognizer*, the Hearer *knows* utterances. These aspects of the domain allow us to conclude that it is the Speaker-agent that occupies the first row of the domain formulation table, and that the Hearer-agent will occupy the second. For convenience, we have labelled the agents with the coordinates of the box they occupy.

The domain can be implemented in at least four different ways, corresponding to the four different possible combinations of Speaker and Hearer, as represented in the domain formulation table. The four possible implementations are enumerated in table 3.3. The first column of table 3.2 represents a system where both members of the speaker–hearer pair *know* their *explananda*; but due to the nature of the domain itself, these *explananda* will be different. Likewise for column two, where the *explananda* are *unknown.*

Speaker–Hearer Duality is a feature of the domain which gives rise to the *Shared Information Constraint*, which suggests that there are two reasonable ways to assign grid

	Inference Direction	Speaker		Hearer	
		knows	uses	knows	uses
I	utt ⇒ bel	bel	abduction	utt	prediction
II	bel ⇒ utt	bel	prediction	utt	abduction

Table 3.4: Speaker–Hearer Duality

positions to speaker and hearer, and consequently, that there are two sensible implementation strategies.

3.3.2 The Shared–Information Constraint

We have already argued that a certain (probably large) percentage of the information available to hearer and speaker must be mutual to them both for successful communication. We suggest now that this places a useful constraint upon domain axiomatization, and gives us a partial answer to our implementation question: for the speaker and hearer to share knowledge, their worlds should be axiomatized *the same way.* In particular, given a set of principles of communication which express ('causal') relations between beliefs and utterances, the Speaker and Hearer should adopt the same view of this causality. This means that, for either of the axiomatizations presented, the two members of the speaker-hearer pair will use different inference mechanisms, *viz.* abductive or predictive reasoning. (Refer to table 3.4). We will call this useful domain–formulation constraint the *The Shared Information Constraint.*

Observe that there are (at least) two essentially distinct approaches to axiomatizing the speaker-hearer pair's communication domain. These correspond to what we have referred to loosely as the "directions of inference", and are labelled with roman numerals in table 3.2. Note that in both cases, the Speaker is performing *Design*, while the Hearer is involved in *Recognition*; it is their *explananda* –along with the inference strategies they adopt– that vary depending upon their grid positions.

In addition to the *Shared Information Constraint,* there are independent concerns which also motivate and which may constrain the implementation methodology. These are addressed in the following sections.

3.3.3 Alternative Implementation Strategies

Having accepted the argument for mutually represented information to be compelling enough to constrain the formulation of the domain, there are still two alternatives. Any domain is likely to admit of this kind of 'vagueness', which is not unlike the problem of choosing an algorithm in a conventional programming language.

Case I

Choosing the axiomatisation of case I means the hearer agent uses prediction, while the speaker agent uses abduction, and that the principles of communication will be of the

following form:[6]

$$H_I = \left\{ \begin{array}{l} principle_1 \\ principle_2 \\ \vdots \\ principle_m \\ utt(X,Y) \end{array} \right\}$$

$$F_I = \left\{ \begin{array}{l} principle_1 \wedge utt(\alpha,\omega) \Rightarrow bel(\alpha, B_{11}) \wedge bel(\alpha, B_{12}) \wedge \cdots \wedge bel(\alpha, B_{1b_1}) \\ principle_2 \wedge utt(\alpha,\omega) \Rightarrow bel(\alpha, B_{21}) \wedge bel(\alpha, B_{22}) \wedge \cdots \wedge bel(\alpha, B_{2b_2}) \\ \vdots \\ principle_m \wedge utt(\alpha,\omega) \Rightarrow bel(\alpha, B_{m1}) \wedge bel(\alpha, B_{m2}) \wedge \cdots \wedge bel(\alpha, B_{mb_m}) \end{array} \right\}$$

In adopting the predictive approach for the hearer, we consider the facts F to consist in the utterances themselves and all other information regarded as true; thus the utterances are the observations which are to be *explained,* or 'diagnosed'. H is *inter alia*[7] the default representation of the principles of communication, *viz,* the normality assumptions. For instance, a speaker is normally sincere, thereby believing what she says. We are prepared to accept sincerity as 'normal' (equation 3.1), and as a component in the diagnosis, as in equation 3.2.

$$\begin{array}{c} sincere(Speaker,\omega) \\ lying(Speaker,\omega) \end{array} \tag{3.1}$$

$$sincere(S,\omega) \wedge utt(S,\omega) \Rightarrow bel(S,\omega) \wedge relevant(S,\omega) \tag{3.2}$$

Sarcasm, misdirection, and outright lying are also possible explanations of the observations, and may enter into the Hearer's recognition process as in equation 3.3.[8]

$$lying(S,\omega) \wedge utt(S,\omega) \Rightarrow \neg bel(S,\omega) \wedge \neg bel(S, \neg bel(Hearer,\omega)) \tag{3.3}$$

This is perhaps a sceptical view of human communication, but lying *is* a well-established human trait. It is only reasonable to presume that our artificial interlocutors will someday fall prey to unscrupulous users unless forewarned of our propensity to mislead![9]

The Speaker uses the default representation of the principles of communication, along with her beliefs, to abduce utterances which fulfil her design objectives.

---

[6]Some of these *facts* actually function as hypotheses in our implementation; this distinction is unimportant here...

[7]Both the theory and the implementation posit other elements which also add default rules, but which can be ignored for our purposes here

[8]See [16] for a description of the predicates involved.

[9]The system may not be able to *predict* any particular belief component of a mental state, even though it may be able to *explain* this component. In this way, the **UM** can entertain competing models of the Speaker's mental state.

Case II

Choosing the axiomatisation of case II means the hearer agent uses abduction, while the speaker agent uses prediction, and that the principles of communication will be represented in the following form:

$$H_{II} = \left\{ \begin{array}{l} principle_1 \\ principle_2 \\ \vdots \\ principle_m \\ bel(X,Y) \end{array} \right\}$$

$$F_{II} = \left\{ \begin{array}{l} bel(\alpha, B_{11}) \wedge bel(\alpha, B_{12}) \wedge \cdots \wedge bel(\alpha, B_{1b_1}) \wedge principle_1 \Rightarrow utt(\alpha,\omega) \\ bel(\alpha, B_{21}) \wedge bel(\alpha, B_{22}) \wedge \cdots \wedge bel(\alpha, B_{2b_2}) \wedge principle_2 \Rightarrow utt(\alpha,\omega) \\ \vdots \\ bel(\alpha, B_{m1}) \wedge bel(\alpha, B_{m2}) \wedge \cdots \wedge bel(\alpha, B_{mb_m}) \wedge principle_m \Rightarrow utt(\alpha,\omega) \end{array} \right\}$$

The principles of communication can be regarded here as possible hypotheses which would be acceptable as explanations of the observations. Stated in diagnostic terms, the principles would be the possible causes of the observed symptoms. Thus, in the presence of a conjectural *intention*[10] on the part of the speaker to communicate, one explanation of an observed utterance is based on conjectured sincerity.

$$bel(\text{Speaker},\omega) \wedge relevant(Speaker,\omega) \wedge sincere(\text{Speaker},\omega) \Rightarrow utt(\text{Speaker},\omega) \quad (3.4)$$

The facts for the Speaker are her beliefs, which are to be explained with those of the default principles which are consistent.

The reader should note here that there is a formulation and implementation of **Theorist** which allows for both abduction and prediction to be performed within the same framework, on a single database. This architecture, shown in Figure 3.3, is suited to implementing the communications domain of the Speaker and Hearer agents described in this chapter.[11] Figure 3.3 depicts the implementation alternative described in this chapter as Case I.

3.4 Summary

Finding enough constraints in a domain to uniquely define its default axiomatisation is not usually possible. Default implementations can be classified along (at least) two dimensions: the assumption and *explananda* status dimensions, which we have represented as the rows and colums of the domain–formulation grid. The domain formulation task can be superficially regarded as one of finding how the domain fits into the grid's representation framework. The example of the simple communication domain demonstrates this process, and is particularly illustrative for a number of reasons:

[10] *I.e.*, in the presence of some belief–predicated term or terms expressing the Speaker's belief that her utterance is relevant, etc.

[11] However, as noted elsewhere in this thesis, I have implemented only the Hearer's side of the conversation.

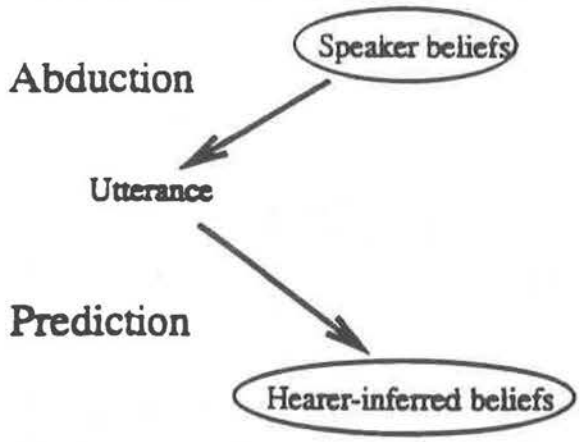

Figure 3.3: Theorist Architecture for Abduction and Prediction

- it exhibits aspects along both dimensions
- it is characteristic of the plurality of implementation strategies of which many domains will admit
- it brings to light a general constraining principle, the *shared–information constraint*

Consideration of the *Shared Information Constraint* led us to an implementation that allows the Speaker and Hearer agents to share knowledge.

A general observation stemming from this work and which may lead to further research is preliminary evidence for the claim that extant logical databases designed for planning applications might be pressed into service as databases for recognition applications, and vice-versa. An important question to be answered in this regard concerns the amount of attention that must be devoted during the design of the database to ensure that this dual usage will be possible.[12]

The cognitive ramifications may be of interest as well. It is noteworthy that in the context of the work presented in this paper, there is no way to decide externally, whether an agent engaged in communication is employing a particular axiomatisation (*e.g.*, case I or case II, etc.), or using a particular inference strategy (*i.e.*, abduction or deduction).

[12]Similar considerations motivate the concept of code reuse in software engineering.

Chapter 4

Implementation

The beliefs of today may count as true today, if they carry us along the stream; but tomorrow they will be false, and must be replaced by new beliefs to meet the new situation.

—Russell on Bergson's Finalism.

The implementation is presented and discussed in this section. It is written entirely in the **Theorist** language. The code portions are presented in distinct units, loosely corresponding to the categories identified in the meta-schema presented earlier; in some cases, the code and sample output have been edited for readability. The unabridged code for the entire implementation is reproduced in Appendix A.[1]

After an introduction to the implementation language, I return to discuss presuppositions and the principles of communication, with an eye to isolating their roles in the current project.

4.1 Implementation Language

The underlying representation language is that of **Theorist,** as described in § 2.5.1. I define rules to represent various types of information, as described throughout this thesis, and particularly as distinguished in chapter 2.[2] The categories of interest are:

- The maxims of the cooperative principle
- Presupposition generating rules (from lexical categories)
- Implicature generating rules
- *Ad-hoc* belief support functions

The following can be considered a *meta-schema* of the predictive version of the implementation. Figure 4.1 describes the form in which the maxims are to be captured. The interpretation I intend for the syntactic elements are as follows:

- utt(α, ω): The agent α 'utters' the statement ω.

[1] Some of this work appeared in *From Utterance to Belief,* by Csinger & Poole [17].

[2] Each of the following categories are represented by an inference path in Figure 3.1.

- bel(α, β): The agent α 'believes' the statement β.
- imp(α, ι): The agent α 'implicates' the statement ι.
- pre(α, π): The agent α 'presupposes' the statement π.

As for the meanings of the quoted terms, I would like to leave their definitions as pre-theoretic as possible. Hadley has surveyed [24] the use of *belief* in the field of AI, and has concluded that it is unclear to what extent the various theories are taken by their proponents to be *true* theories, or realistic cognitive models. He also adds that the 'syntactic approach' underlies the others to varying degrees. With this in mind, and with the conviction that a realistic account of (human) cognition need not necessarily be logical in any sense, I do not wish to go beyond a syntactic characterization of the current model. In leaving the definitions as 'pre-theoretic' as possible, I mean to avoid imposing either a semantics or a claim to psychological validity. If history continues as it has in recent years, the lifespans of such claims are not likely to be long.

Thus, I can say that *utterances* are context-situated[3], that $utt(\alpha, \omega)$ means the agent α expresses a statement ω. The information content of ω is its propositional content, augmented with the inferences sanctioned by both the rules of the cooperative principle, and the context embodied in the beliefs of the agent and those of his interlocutors.

An agent α *believes* the information expressed by β just in case the quantity $bel(\alpha, \beta)$ holds true. As noted above, I hold fast to the syntactic view, by which device two expressions β_1 and β_2 are different, even if they can be considered synonymous under some semantically defined operation. Thus, I leave open the question of whether an agent who believes *Mary has a brother* also believes *Mary has a male sibling.* As far as my implementation goes, agents will not perceive such synonymies unless presented with an explicit rule to identify them.

An agent α *implicates* an expression ι just in case the quantity $imp(\alpha, \iota)$ holds true. This happens when an inference is sanctioned by the line connecting *utterance* to *implicature* in Figure 3.1. ι can not be both implicated in this sense and *uttered* as described above. *I.e.,*

$$\forall_{\omega,\iota} \mathrm{utt}(\alpha, \omega) \wedge \mathrm{imp}(\alpha, \iota) \wedge \omega \neq \iota$$

An agent α *presupposes* an expression π just in case the quantity $pre(\alpha, \pi)$ holds true. This happens when an inference is sanctioned by any line terminating in *presupposition* in Figure 3.1. π can not be presupposed in this sense if it is either *uttered* or *implicated* as described above. *I.e.,*

$$\forall_{\omega,\iota,\pi}(\mathrm{utt}(\alpha, \omega) \wedge \mathrm{pre}(\alpha, \pi) \wedge \omega \neq \pi) \vee (\mathrm{imp}(\alpha, \iota) \wedge \mathrm{pre}(\alpha, \pi) \wedge \iota \neq \pi)$$

In Figure 4.1, the $B_{i,j}$ are the beliefs adduced to capture the normative strengths of the maxims as discussed in the relevant sections of this thesis.

In Figure 4.2, the ι's are derived from the forms of the ω's; this places constraints on the ι's sufficient to guarantee, for instance, that $\omega \neq \iota$.

In Figure 4.3, the π's are derived from the forms of the ω's; this places constraints on the π's sufficient to guarantee, for instance, that $\omega \neq \pi$. In addition to the default rules of Figures 4.1, 4.2, and 4.3, the rules of 4.1 and 4.2 are needed to derive beliefs describing

[3]Which is to say little more than that the theory I am constructing is a *pragmatic* one.

$$
\begin{aligned}
&default \quad principle_1 : utt(\alpha,\omega) \Rightarrow bel(\alpha, B_{11}), bel(\alpha, B_{12}), \cdots, bel(\alpha, B_{1b_1}).\\
&default \quad principle_2 : utt(\alpha,\omega) \Rightarrow bel(\alpha, B_{21}), bel(\alpha, B_{22}), \cdots, bel(\alpha, B_{2b_2}).\\
&\quad\vdots\\
&default \quad principle_m : utt(\alpha,\omega) \Rightarrow bel(\alpha, B_{m1}), bel(\alpha, B_{m2}), \cdots, bel(\alpha, B_{mb_m})
\end{aligned}
$$

Figure 4.1: Principles of Communications

$$
\begin{aligned}
&default \quad implicature_1 : utt(\alpha,\omega), principle_{y_1} \Rightarrow imp(\alpha, \iota_{11}), imp(\alpha, \iota_{12}), \cdots, imp(\alpha, \iota_{1i_1})\\
&default \quad implicature_2 : utt(\alpha,\omega), principle_{y_2} \Rightarrow imp(\alpha, \iota_{21}), imp(\alpha, \iota_{22}), \cdots, imp(\alpha, \iota_{2i_2})\\
&\quad\vdots\\
&default \quad implicature_p : utt(\alpha,\omega), principle_{y_p} \Rightarrow imp(\alpha, \iota_{p1}), imp(\alpha, \iota_{p2}), \cdots, imp(\alpha, \iota_{pi_p})
\end{aligned}
$$

Figure 4.2: Implicature Generators

$$
\begin{aligned}
&default \quad presupposition_1 : utt(\alpha,\omega), principle_{x_1} \Rightarrow pre(\alpha, \pi_{11}), pre(\alpha, \pi_{12}), \cdots, pre(\alpha, \pi_{1r_1})\\
&default \quad presupposition_2 : utt(\alpha,\omega), principle_{x_2} \Rightarrow pre(\alpha, \pi_{21}), pre(\alpha, \pi_{22}), \cdots, pre(\alpha, \pi_{2r_2})\\
&\quad\vdots\\
&default \quad presupposition_s : utt(\alpha,\omega), principle_{x_s} \Rightarrow pre(\alpha, \pi_{s1}), pre(\alpha, \pi_{s2}), \cdots, pre(\alpha, \pi_{sr_s})
\end{aligned}
$$

Figure 4.3: Presupposition Schemas/Triggers

$$
\begin{array}{ll}
default & rationality_1 : bel(\alpha, B_1) \wedge principle_{y_1} \Rightarrow bel(\alpha, B_{11}) \wedge bel(\alpha, B_{12}) \wedge \cdots \wedge bel(\alpha, B_{1r_1}) \\
default & rationality_2 : bel(\alpha, B_2) \wedge principle_{y_2} \Rightarrow bel(\alpha, B_{21}) \wedge bel(\alpha, B_{22}) \wedge \cdots \wedge bel(\alpha, B_{2r_2}) \\
\vdots & \\
default & rationality_s : bel(\alpha, B_s) \wedge principle_{y_s} \Rightarrow bel(\alpha, B_{s1}) \wedge bel(\alpha, B_{s2}) \wedge \cdots \wedge bel(\alpha, B_{sr_s})
\end{array}
$$

Figure 4.4: Rationality Constraint Schema

the mental state of the speaker (or rather, that of the system which models the mental state of the speaker).

$$default \text{ bel\_imp}(\alpha, \iota) : imp(\alpha, \iota) \Rightarrow bel(\alpha, \iota). \tag{4.1}$$

$$default \text{ bel\_pre}(\alpha, \pi) : pre(\alpha, \pi) \Rightarrow bel(\alpha, \pi). \tag{4.2}$$

Rationality conditions can also be implemented as default rules, representing a set of normative constraints which exhibit the desirable behavior of defeasibility, thus relaxing the traditional requirements of closure and consistency.[4] The rationality (or introspection) schema cannot be implemented directly as shown in Figure 4.4 without some consideration of the underlying control mechanisms. See § 4.5 for details.

Different types of knowledge can be implemented either as facts or as defaults in the logic, depending upon their epistemic status as perceived by the implementor. I have adopted the view that all beliefs are defeasible, as suggested by Shoham[51].

4.2 Principles

Others before me have felt free to implement and reformulate the Gricean Maxims, picking and choosing from among them as they saw fit. I see no reason why I should not indulge in a similar practice, with the accompanying explanations.

I retain of Grice the reasonable working hypothesis that communication is governed by a set of *principles* (which Grice calls his 'maxims'), which would –if completely explicated– provide explanations for natural language utterances. I do not make any claim regarding the number of these governing principles, and will refer instead to the set which contains them, even though its cardinality is unknown.

It is these principles of communication which I implement in this thesis. The relationship between 'my principles' and 'Grice's maxims' is summed up by observing that Grice restricted himself to 'cooperative' forms of communication. The principles I have in mind seek to capture normal [human] communications in a broader normative sense. In particular, different kinds of misleading are normal, rational communicative pursuits, and the theory should be able to represent these. See Figure 4.5. It is worthwhile to my

[4]Consistency remains a criterion of rationality in the implementation I present, but in the default theoretical as opposed to the traditional, monotonic sense. Nonetheless, I do not wish to claim that consistency is in any sense a property of rationality; I know of many empirical counterexamples to such a claim!

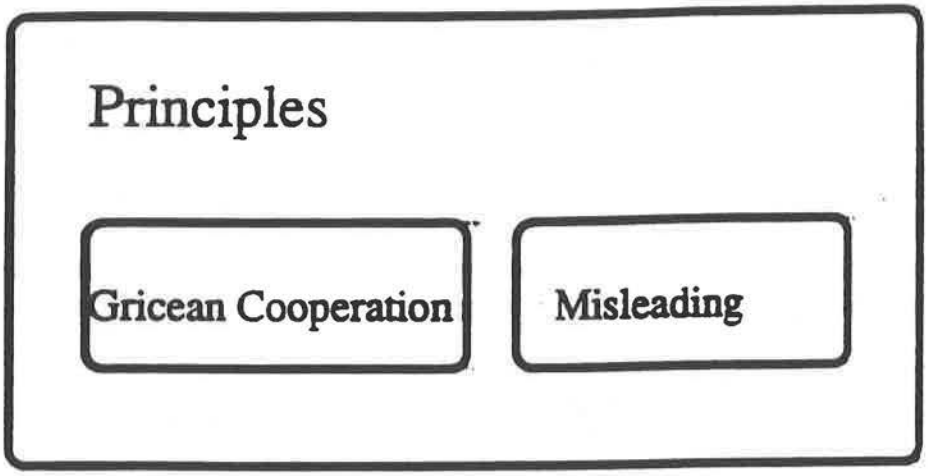

Figure 4.5: Principles and Grice's Maxims

project to bear in mind throughout, the essential defeasibility of any of the principles. All of their exhortations should be prefixed with something along the lines of 'in the absence of any contradictory information...,' or more significantly, perhaps: 'by default...' Thus, while the principles are the expression of the *norms* of human communication, they give way to other, *ab*normal modes of communication, which I lump under the blanket term, *misleading*, to distinguish them from the cooperative mode. (These intuitions are nicely elaborated by others, *e.g.*: van Frassen [55, 52] and Lewis [37]. As promised throughout this thesis, the principles of cooperative communication have been captured in the **Theorist** language, and the resulting implementation is presented in Appendix A. These rules are the simplest that could plausibly account for the inferences involved. Their interactions with the rules expressing presupposition and implicature are described in the upcoming sections dealing with those rules.[5]

The Maxim of Quality is a sincerity condition, the formulation of which follows, and is consistent with Searle's account of Speech Acts [50, p18]: *a Speaker believes what she says.* I will call this the *Principle of Sincerity.*

Quantity is the idea that a speaker should utter the most specific statement of what she wishes to communicate. A reasonable—but by no means exhaustive—formulation of this is that *when a speaker utters a disjunction, he does so because no other natural language connective is expressive of the 'tentativeness' of his belief in either of the disjuncts.* This rule thus sanctions the derivation of the clausal quantity implicatures as per Gazdar [20] and Mercer [39]. This will be the basis of my *Principle of Disjunction.*

[5]Compliance with the principles is *normative*: deviations in non-ideal agents are to be expected, and theories founded on principles of rationality might be pressed into service as lie-detectors of sorts, if not truth-detectors. As described earlier in this thesis, Grice categorized his Cooperative Principle into a number of maxims which were intended to explain natural language communication between cooperating agents. In the discussion which follows, I refer to these categories only because they are a good starting point; I am not committed to a "Gricean" theory, in any deeper sense.

Principle 1 (Sincerity) *A Speaker believes what she says.*

Principle 2 (Disjunction) *A Speaker may believe any of the disjuncts in her utterance.*

Principle 3 (Relevance) *A Speaker believes the hearer does not* a priori *believe her utterance.*

Principle 4 (Sarcasm) *A Speaker does not believe her utterance and*

- *believes the hearer does not believe the utterance*
- *believes the hearer believes that she does not believe her utterance.*

Figure 4.6: Some of the Principles of Communications

Relevance is tricky. I suggest that anyone who can completely formulate this one in *any* kind of logic will have solved most—if not all—of the problems of Artificial Intelligence!! As a first attempt, I might expect the speaker to *utter only what the speaker believes the hearer does not already know.* Groenendick and Stockhof [23] have referred to this as a principle of *informativeness.* This becomes my *Principle of Relevance.*[6]

Perspicuity is too vague a concept to admit of an obvious representation within the current framework, and I will leave it for future work.

Sarcasm, though not one of the 'original' maxims, can be captured simply along the following lines. A speaker is sarcastic when *the speaker 1) does not believe her utterance, 2) believes that the hearer does not believe the utterance, and 3) believes that the hearer believes that the speaker does not believe the utterance.* These conditions mark my *Principle of Sarcasm.*

The principles discussed above are summarized in Figure 4.6, and their **Theorist** implementations appear in Appendix A.

Example 4–1: Beliefs derived given:

```
fact utt(dave,not property(john, regret, jumping)).

bel(dave,not property(john,regret,jumped))
sincere(dave,not property(john,regret,jumped))

bel(dave,not bel(hearer,not property(john,regret,jumped)))
sarcastic(dave,not property(john,regret,jumped))

bel(dave,bel(hearer,bel(dave,not bel(hearer,not property(john,regret,jumped)))))
sarcastic(dave,not property(john,regret,jumped))
```

■

Note that a speaker can not be simultaneously sincere and sarcastic with respect to a particular utterance. Whereas a conventional logical approach would derive a contradiction,

[6]The case where a speaker utters ω even though she believes that the hearer already believes ω, is not covered by this principle of relevance, but would be explainable via a principle of *confirmation.*

or not conclude anything, the mechanics of default reasoning derive the consequences of assuming both sincerity and sarcasm, with mutually inconsistent beliefs residing in separate extensions of the resulting theory. For example, given the utterance by *Dave* that he regrets that *John* jumped, **Theorist** derives the formulae of example 4–1.[7] See § 2.5.2 for a discussion of how one extension might be 'preferred' over another.

4.3 Presupposition

Mercer [39] shows how to represent a number of presuppositional schemas in Reiter's formalism for default logic. These schemas correspond largely to the fragments of **Theorist** code presented in this section. **Theorist** provides a useable implementation, and I have reified over properties to allow for a first-order representation.

4.3.1 Criterial and Non-criterial Properties

Mercer describes his schema for non-criterial properties in terms of the meaning-inheritance hierarchy of a lexeme. The criterial properties of a lexeme are those which define the terminal branches of the hierarchy, *e.g.*, a bachelor is unmarried. Non-criterial properties are those which define the other levels of the hierarchy, *e.g.*, bachelors are [generally] male, and adult. Mercer says of this category of presupposition that it is a "type of lexical presupposition which is based on the deciding criterion of a negated lexeme's meaning" [39, p76].

Example 4–2:

Sentence 4–1: My cousin is not a bachelor
Sentence 4–2: The speaker's cousin is male
Sentence 4–3: The speaker's cousin is adult ■

Mercer's example is reproduced as example 4–2. Given the utterance of sentence 4–1, the presuppositions of sentences 4–2 and 4–3 can be derived. The non-criterial presupposition schema is implemented as the **Theorist** default rule of Figure 4.7. This rule might be paraphrased as *when a negated lexical item appears in an utterance, and it has non-criterial properties, then if it is consistent to do so, infer that the speaker believes the indicated presupposition.* The non-criterial properties of the lexemes, where applicable, are simply provided as facts in **Theorist.**

Example 4–3: Speaker's Beliefs about Bachelors, given :

```
fact utt(andrew,property(cousin, not bachelor))

Answer is believes(andrew, property(cousin,male))
Theory is [pre_by_nonc(andrew,cousin,bachelor,male)]

Answer is believes(andrew, property(cousin,adult))
Theory is [pre_by_nonc(andrew,cousin,bachelor,adult)]
```

■

[7] Among others which have been omitted here for clarity.

```
% noncriterial presupposition schema:
default pre_by_nonc(S, Object, Propty, Presupposition) :
                (utt(S, property(Object, not Propty)) or
                 imp(S, property(Object, not Propty)))
                and
                nonc(Propty, Presupposition)
                =>
                bel(S, property(Object, Presupposition)).
```

Figure 4.7: Non-criterial default schema

Given the utterance by the agent *Andrew* of *My cousin is not a bachelor,* **Theorist** ascribes the beliefs of example 4–3 to *Andrew.*[8] Note that the antecedent of the presupposition schema contains a conjunct that is a disjunct of an utterance formula and an implicature formula. This is a reflection of the fact that implicatures can themselves sanction presuppositions; this will become clearer in the following section dealing with the implementation of implicatures, and again in § 4.7.

4.3.2 Factive Verbs

Utterances with factive verbs imply the relative clause, whether the verb is negated or not [27].

Example 4–4: Presupposition by Factive Verb, given:

fact utt(andrew, not property(john, regret, came(mary, party)))

Answer is bel(andrew,bel(john,came(mary,party)))
Theory is [pre_by_factive(andrew,john,came(mary,party),regret)] ■

The utterance by *Andrew* that *John regrets that Mary came to the party* entails that (Andrew believes that John believes that)[9] Mary came to the party. The negated form *John does not regret that Mary came to the party* presupposes the same thing. It is with the latter relationship that this implementation is concerned. Example 4–4 gives the presuppositions derived by application of the rule for factives, from the utterance by *Andrew* of *John doesn't regret that Mary came to the party.*

4.4 Implicatures

I restrict myself here first of all to so-called clausal quantity implicatures, and second, to their appearance in disjunctive utterances. Other complex sentences carry similar implicatures (*e.g., if–then* sentences). (See Definition 3 in chapter 2 of this thesis). For instance, when a speaker utters a sentence of the form *A is X or A is Y*, she may mean

[8]Other beliefs are sanctioned as well, deriving from explanations of sincerity, sarcasm, etc., but they have been omitted in the interests of brevity and clarity. See appendix B for unabridged sample sessions with the system.

[9]As noted elsewhere, this work follows Horton in that presuppositions are beliefs of agents.

```
% factive presupposition schema:
default pre_by_factive(Speaker, Subject, Presupposition, Factive) :
(utt(Speaker, not property(Subject, Presupposition, Factive))
                  or
imp(Speaker, not property(Subject, Presupposition, Factive)))
and factive(Factive)
=>
                bel(Speaker, bel(Subject,Presupposition)).
```

Figure 4.8: Factive Verb Presupposition Schema

any of *A is X, A is Y, A is not X, A is not Y.* These are the so-called clausal quantity implicatures, and Mercer assumes their *a-priori* existence in his method for generating the presuppositions of complex sentential forms. It is my intention here to show that they can be accomodated within the theory presented in this thesis, and (equivalently) that they can be produced by the implementation.

The intent in both Mercer's work and in this thesis is that those implicatures which are consistent (mutually and with existing context) will themselves carry presuppositions, and thus sanction additional inferences for the hearer about the mental state of the speaker. The 'survivability' of these potential implicatures is thus a central issue.

Example 4–5: Candidate clausal quantity implicatures given:

Sentence 4–4:	My cousin is a bachelor or a spinster
Sentence 4–5:	My cousin is a bachelor
Sentence 4–6:	My cousin is not a bachelor
Sentence 4–7:	My cousin is a spinster
Sentence 4–8:	My cousin is not a spinster ■

Consider example 4–5. The utterance of sentence 4–4 produces the candidate implicatures of sentences 4–5 through 4–8. In this case, some of the candidates are mutually inconsistent, and thus should be placed in separate extensions of the default theory, for further consideration.

Several obvious choices present themselves for the implementation of the implicature generating rules, with interesting methodological repercussions. Of interest are the following:

1. a single disjunctive default
2. a single conjunctive default
3. separate default rules

Briefly, the first option suggests a default rule of the following form:

$$utt(S, A \cup B) \Rightarrow imp(S, A) \cup imp(S, \neg A) \cup imp(S, B) \cup imp(S, \neg B)$$

With reference to example 4–5, this approach can be easily dismissed, for it is *too weak*; it allows the survival in a single extension of mutually inconsistent candidates, and will

```
% Clausal quantity implicature generating function, following Gazdar:
default fc(1,S,U1,U2) :
            utt(S, or(U1,U2)) => imp(S,U1).
default fc(2,S,U1,U2) :
            utt(S, or(U1,U2)) => imp(S,U2).
default fc(3,S,U1,U2) :
            utt(S, or(U1,U2)) => imp(S,not U1).
default fc(4,S,U1,U2) :
            utt(S, or(U1,U2)) => imp(S,not U2).
```

Figure 4.9: Implicature-generating schema

subsequently sanction the prediction of invalid presuppositions, resulting in a mental model of the speaker that is patently incorrect.

The second option requires a default rule of the following form:

$$utt(S, A \cup B) \Rightarrow imp(S, A) \cap imp(S, \neg A) \cap imp(S, B) \cap imp(S, \neg B)$$

This approach is *too strong*; if any of the candidate implicatures are inconsistent (with context of with another candidate), then *none* of them will be predicted. This is because the conjunction requires that all of the candidates be true in some (single) extension of the default theory.

The last choice is a set of four default rules, one for each of the candidate implicatures in the disjunctive environment. This has the intended effect of letting only those candidates survive that are consistent with established context, while maintaining alternate possibilities.

Figure 4.9 shows a possible implementation resulting from the third approach discussed above.

The preceding discussion has been left at a deliberately intuitive level, as nothing would be gained from additional formality. The intent has been to give a justification of the approach taken to the implementation of the implicature generating rules, and to provide a feeling for some of the default-logic programming issues that arise in practice.

4.5 Rationality

Some aspects of the rationality conditions could not be implemented in **Theorist** without attention to the underlying control mechanism. The expressive power of **Theorist** is gained at the expense of not being able to guarantee the computability of an expression. In particular, some formulae which intuitively capture the obvious properties of introspection are patently left-recursive, with the result that pure **Theorist** will not terminate in evaluating these expressions.

To alleviate these restrictions, a simple depth-bound has been imposed upon the mechanics of the theorem-prover. The repercussions for the implementation are that left-recursive formulae can be evaluated up to the depth-bound. The theory itself is compromised in that completeness and soundness can be no longer simultaneously ensured.

```
% We are informed of knowledge that is mutually known in the community:
default aware(Agent,A) :
      mutual(A) => bel(Agent,A).

% If we believe the antecedent of a rule, we believe its consequent:
default implicit(Agent,A,C) :
      mutual(=>(A,C)) and bel(Agent,A) => bel(Agent,C).

% If we believe a list, we believe the items in the list:
default conjunct(Agent,List,X) :
      bel(Agent,List) and member(X,List) => bel(Agent,X).

% Positive Introspection (patently left-recursive):
default pos_int(Agent,B) :
      bel(Agent,B) => bel(Agent,bel(Agent,B)).
```

Figure 4.10: Rationality Constraints

However, all derivations in this implementation are unaffected by this loss.[10] A depth-bound is a natural kind of restriction to place upon an inference mechanism, reflecting the finiteness of the agent concerned [10].

Figure 4.10 are some default rules that express likely conditions on rationality or introspection. They correspond to previous efforts by other researchers as related in earlier chapters of this thesis, and alleviate the problem of logical omniscience by relaxing the well-formedness criteria to one of default, rather than classical logical consistency.

4.6 Other Aspects

Other kinds of information are also required by the theory, and must be represented in the implementation. In particular, world information, lexical information, etc., as discussed earlier, must be provided for. Refer to Appendix A.1.5 for details.

4.7 Cancellation and Multiple Extensions

Mercer has provided an explication of how default logic might be employed to represent and derive the presuppositions of natural language utterances, going as far as to show how this might be done for complex sentences such as disjunctions. His technique is to avoid multiple extension theories wherever possible, as there is no clear semantics for theories of this type, and only a hazy ontology. This is a general problem with default reasoning, and most practitioners have sought to avoid it, rather than solve it.

[10] A version of **Theorist** which employs iterative deepening search strategies is under development. This version will be both sound and complete, and will exhibit all the desireable features of the depth-bounded implementation.

Although Mercer urges that *in the case of multiple extension theories,* the actual presuppositions of a complex utterance are those which are in all extensions, he is unhappy with his definition because he cannot provide a clear interpretation of "membership in all extensions." I, on the other hand, have argued in this thesis that the extensions of a default theory can be regarded as mere technical components of a system, insofar as they serve to expedite the process of presupposition generation, and I have noted the correspondence of this claim to Gazdar's notion of *pre*-supposition. I am now prepared to go a little farther.

When the Speaker utters **Jack is a bachelor or a spinster**, the (sceptical) criterion of membership-in-all-extensions permits only the derivation of the Speaker-belief that **Jack is an adult**. In particular, the system is unable on these grounds to decide the sex of Jack. But if the theory also includes a default rule to the effect that people with the name Jack are of the male sex, then there is what might be thought of as reinforcing evidence for the Speaker-belief that Jack is a male. This extra information can also be regarded as a new counter-argument against the Speaker-belief that Jack is a female. It is this intuition that I would like to promote as the basis for theory preference (see § 2.5.2).

Chapter 5

Conclusion

> *...why may we not say, that all* Automata *(Engines that move themselves by springs and wheeles as doth a watch) have an artificiall life?*
>
> —Thomas Hobbes, *Leviathan*

> *Let us likewise beware of believing the universe is a machine; it is certainly not constructed so as to perform some operation, we do it far too great honour with the word 'machine'.*
>
> —Nietzsche, *The Gay Science.*[1]

5.1 Contribution

This thesis has made contributions in several areas.

- A principled theory communication has been developed, with particular attention to its application in the field of user-modelling
- Mercer's [39] theory of presupposition has been extended to include beliefs of interlocutors [27], and has additionally been implemented in the **Theorist** [46] framework for default reasoning
- The theory of communication has also been implemented in the same framework for default reasoning, allowing derivation of implicatures and presuppositions
- The theory and implementation support the derivation of users' beliefs from their utterances, thereby demonstrating the application of default reasoning theory and practice to user-modelling
- Issues of default logic programming have been resolved, with resulting contributions to that body of knowledge

[1] From the introduction to *Thus Spoke Zarathustra*, p17. Translated by R.J. Hollingdale, Penguin Books, 1969, New York, NY.

- The theory and implementation allow representation of alternate interpretations of the discourse

5.2 Problems

5.2.1 Multiple Extensions

The astute reader of this thesis will have noticed an apparent contradiction, which I have left unresolved to this point. The weakness I to which I refer concerns the thorny issue of multiple extensions, and their differing interpretations within my system in the context of presupposition generation and of utterance-meaning.

I have suggested a purely syntactic and ontologically agnostic view of multiple extensions with regard to their role in presupposition-generation (§ 4.7), while with regard to the application of the principles of communication, I have suggested that a multiplicity of extensions has significant representational importance (§ 4.2, § 2.2.1).

I have been admittedly opportunistic, and a complete resolution of this issue will not disappear until an adequate basis for theory preference is established. I have suggested how this might be accomplished within an ontology that is purely syntactic (§ 2.5.2), and hope to make some progress in this area. The syntactic account resolves the problem described, although the implementation presented in this thesis is not yet able to make use of these observations.

5.2.2 Goals, Plans and Desires

Though I am not yet ready to recant my earlier view that beliefs are enough to represent mental states of interlocutors, I now admit that there are immediate advantages to augmenting the representational language to include primitives for such things as goals and desires.

A user-modeller, for instance, might profit from being able to reason about the user's goals.

5.3 Further work

There are two obvious directions in which to take further work. As noted throughout this thesis, I have systematically avoided trying to account for the *goals* and *desires* of agents represented with this system. My reasons for this are quite practical. Such an effort would have taken me to the outer limits of pragmatics, where I would at best have been on shaky ground. I would then have had to take into account the Speaker's point of view as well, and this would not serve in the development of a User-modeller, for which a completely Hearer-based view is adequate. Of course, none of these disclaimers prevent future expansion of this work to eventually encompass goals and desires of both Speaker and Hearer; the methodology and the reasoning framework employed were chosen to assure that such future efforts would remain consistent with what has already been presented here. Thus, one avenue for future work is the development of principles of pragmatics, to be represented in a default reasoning framework. The search for these

principles would be hampered by lack of any underlying theory, and such an effort should probably be delayed until cognitive science has more to offer.

Another —and I think better— direction to take would be to probe further into the mechanism of the default reasoning framework itself. The current implementation is plagued by the well–known problem of multiple extensions, and any enhancement of the system to cover goals or desires would continue to suffer from these same problems. The still unresolved difficulties of preference in multiple extension theories will continue to be a major impediment to the productive application of default reasoning.

Interesting questions arise when the issues discussed in this thesis are transposed into a communication environment that supports some modality other than natural language. In particular, one wonders how to design an *artificial* language to best support the task of a user-modeller, as described herein. Theoretical issues relating to the semantics of these putative graphical languages arise, as well as practical implementation questions; work is being undertaken on both these fronts at the University of British Columbia.

Bibliography

[1] James Allen. *Natural Language Understanding.* The Benjamin/Cummings Publishing Company, Inc., Menlo Park, CA, 1987.

[2] James Allen and Raymond Perrault. Analyzing intention in utterances. *Artificial Intelligence*, 15:143–178, 1980.

[3] Douglas E. Appelt and Kurt Konolige. A nonmonotonic logic for reasoning about speech acts and belief revision. In *2nd International Workshop on Non-Monotonic Reasoning*, pages 164–176, 1988.

[4] J. L. Austin. *How to do Things with Words.* Oxford University Press, 1962.

[5] Kent Bach and Robert M. Harnish. *Linguistic Communication and Speech Acts.* MIT Press, Cambridge, MA, 1979.

[6] John Batali. Computational introspection. A.I. Memo 701, Artificial Intelligence Laboratory, M.I.T., 1983.

[7] Gerhard Brewka. Preferred subtheories: An extended logical framework for logical reasoning. *IJCAI*, 1989.

[8] Noel Burton-Roberts. *The Limits to Debate: A revised theory of semantic presupposition.* Cambridge University Press, 1989.

[9] E. Charniak. Inference and knowledge i. In E. Charniak and Y. Wilks, editors, *Computational Semantics.* North-Holland, 1976.

[10] Christopher Cherniak. *Minimal Rationality.* MIT Press: A Bradford Book, Cambridge, MA, 1986.

[11] Philip R. Cohen and Hector J. Levesque. Speech acts and rationality. *CSLI.*

[12] Philip R. Cohen and Raymond Perrault. Elements of a plan-based theory of speech-acts. *Cognitive Science*, 3:177–212, 1979.

[13] A. Csinger, H.da Costa, B.Forghani, and D.A.Lowther. Increasing cad throughput with a programmable user interface. In *Official Proceedings of the 3rd Int'l IMS '87, SATECH '87, Part I*, Long Beach, CA., 1987.

[14] A. Csinger, H. da Costa, and B. Forghani. A general-purpose programmable command decoder. In *IEEE Proceedings, Conference Compint*, pages 139–41, November 1987.

[15] Andrew Csinger. Hypothetically speaking: Default reasoning and discourse structure. presented at IRIS/Precarn Conference, June 1990.

[16] Andrew Csinger. Implementing a theory of communications in a default reasoning framework. Master's thesis, University of British Columbia, 1990.

[17] Andrew Csinger and David Poole. From utterance to belief: Default reasoning in user-modelling. In *Proceedings of the Conference for Knowledge Based Computing Systems, KBCS-89*, pages 408–419, Bombay, India., December 1989.

[18] David W. Etherington. Formalizing non-monotonic reasoning systems. Technical Report 1, University of British Columbia, Vancouver, Canada, V6T 1W5, 1983.

[19] Gottlob Frege. On sense and reference. In P. Geach and M. Black, editors, *Translations from the Philosophical Writings of Gottlob Frege*, pages 55–78. Blackwell, 1892.

[20] Gerald Gazdar. *Pragmatics: Implicature, Presupposition and Logical Form.* Academic Press, 1979.

[21] Hector Geffner. Default reasoning, minimality and coherence. In *KR89*, page 137, Toronto, Canada, May 1989.

[22] H.P. Grice. Logic and conversation. In P. Cole and J.L. Morgan, editors, *Syntax and Semantics: Speech Acts, vol 3*, pages 47–58. Academic Press, New York, 1975.

[23] Jeroen Groenendijk and Martin Stokhof. A pragmatic analysis of specificity. In Frank Heny, editor, *Ambiguities in Intensional Contexts*, pages 153–190. D. Reidel Company, Dordrecht, Holland, 1980.

[24] Robert F. Hadley. The many uses of 'belief in ai'. CSS-IS TR 03, Centre for Systems Science, Simon Fraser University, Burnaby, B.C., V5A 1S6, 1989.

[25] Hans G. Herzberger. Dimensions of truth. In *Contemporary Research in Philosophical Logic and Linguistic Semantics.* unknown, 1975.

[26] Diane Horton. Incorporating agents' beliefs in a model of presupposition. Technical Report 201, University of Toronto, 1987.

[27] Diane Horton and Graeme Hirst. Presuppositions as beliefs. In *COLING*, 1988.

[28] S.J. Kaplan. Cooperative responses from a portable natural language query system. *Artificial Intelligence*, 19(2):165–88, 1982.

[29] L. Karttunen. Presupposition and linguistic context. *Theoretical Linguistics*, 1:181–194, 1974.

[30] L. Karttunen and S. Peters. Conventional implicature. In C. K. Oh and D. A. Dineen, editors, *Syntax and Semantics.* Academic Press, 1979.

[31] Robert Kass and Tim Finin. Modelling the user in natural language systems. *Computational Linguistics*, 14(3):5, September 1988.

[32] Kurt Konolige. Belief and incompleteness. Technical Report 319, SRI, SRI International, Menlo Park, CA, 1984.

[33] Kurt Konolige. A computational theory of belief introspection. In *IJCAI85*, pages 502–508, 1985.

[34] Robert R. Korfhage. Intelligent information retrieval: Issues in user modelling. Technical Report 85–CSE–9, Dept. of Computer Science and Engineering, Southern Methodist University, Dallas, Texas, May 1985.

[35] Hector J. Levesque. A logic of implicit and explicit belief. *AAAI*, pages 198–202, 1984.

[36] Robert I. Levine. *A Comprehensive Guide to Expert Systems: Turbo-Pascal Edition.* McGraw-Hill, 1988.

[37] D. Lewis. *Convention.* Cambridge University Press, 1969.

[38] William G. Lycan. *Logical Form in Natural Language.* MIT Press, (A Bradford Book), Cambridge, MA, 1984.

[39] Robert Mercer. A default logic approach to the derivation of natural language presuppositions. Technical Report 35, University of British Columbia, October 1987.

[40] Robert Mercer and Richard Rosenberg. Generating corrective answers by computing presuppositions of answers, not of questions. In *Proceedings of the 1984 Conference*, pages 16–19, University of Western Ontario, London, Ontario, May 1984. Canadian Society for Computational Studies of Intelligence.

[41] C. Raymond Perrault. An application of default logic to speech act theory. CSLI 90, Center for the Study of Language and Information, Stanford, CA., 1987.

[42] C. Raymond Perrault and Barbara J. Grosz. Natural language interfaces. *Annual Review of Computer Science*, 1:47–82, 1986.

[43] G.E. Pfaff, editor. *User Interface Management Systems.* Springer–Verlag, Berlin, 1985.

[44] David Poole. On the comparison of theories: Preferring the most specific explanation. In *IJCAI*, volume I, pages 144–147, Los Angeles, CA, August 1985.

[45] David Poole. A logical framework for default reasoning. *Artificial Intelligence*, 36(1):27–47, 1987.

[46] David Poole. Representing knowledge for logic-based diagnosis. In *International Conference on Fifth Generation Computing Systems*, pages 1282–1290, Tokyo, Japan, November 1988.

[47] David Poole. Explanation and prediction: an architecture for default and abductive reasoning. *Computational Intelligence*, 5(2):97–110, 1989.

[48] David Poole. Normality and faults in logic-based diagnosis. In *IJCAI*, pages 1304–1310, Detroit, MI, August 1989.

[49] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1,2):81–132, 1980.

[50] John R. Searle and Daniel Vanderveken. *Foundations of Illocutionary Logic.* Cambridge University Press, 1985.

[51] Yoav Shoham and Yoram Moses. Belief as defeasible knowledge. STAN-CS 1237, Department of Computer Science, Stanford University, Stanford, CA 94305, 1988.

[52] Robert Stalnaker. Review of gazdar's pragmatics. *Language*, 56(4), 1980.

[53] Stephen Stich. *From Folk Psychology to Cognitive Science: The Case Against Belief.* MIT Press (A Bradford Book), Cambridge, MA, 1983.

[54] P. F. Strawson. On referring. *Mind*, 59:320–344, 1950.

[55] Bas C. van Frassen. Incomplete assertion and belnap connectives. In *Contemporary Research in Philosophical Logic and Linguistic Semantics.* unknown, 1975.

Appendix A

Theorist Listings

A.1 Maxims

```
% This version sets out to construct the agents' knowledge bases from
%  an understanding of the Gricean Maxims of Cooperation, and from the
%  utterances of the agents.

%% GRICEAN Quality Analog
% sincerity does not involve TRUE beliefs of the hearer:
default sincere(S, U).

fact sincere(S,U) and
             utt(S, U) => bel(S, U).

%% GRICEAN Quantity Analog
%  the speaker doesn't necessarily believe what he says here; this is
%  subsumed in the sincerity rule:
default quantity(S, U).

fact quantity(S,U) and
             utt(S, U) => bel(S, not bel(hearer, U)).
             % don't say what you know hearer knows


% Sarcasm predication
default sarcastic(S, U).

fact sarcastic(S,U) and
             utt(S, U)
             =>
             bel(S, not bel(hearer, U)) and
             bel(S, bel(hearer, bel(S, not bel(hearer, U)))) and
             not bel(S, U).
```

A.2 Presupposition

```
% PRESUPPOSITIONAL ANALYSES:
% default rules to enable presuppositions under negation:
%  noncriterial presupposition schema:
default pre_by_nonc(S, Object, Propty, Presupposition) :
                (utt(S, property(Object, not Propty)) or
                 imp(S, property(Object, not Propty)))
                and
                nonc(Propty, Presupposition)
                =>
                bel(S, property(Object, Presupposition)).

% factive presupposition schema:
%  what we really want in the following is the narrow-scope
%  negation of the factive verb, but we adopt the wide scope
%  representation for convenience.
default pre_by_factive(Speaker,Subject,Presupposition,Factive) :
(utt(Speaker, not property(Subject, Factive, Presupposition))
                  or
imp(Speaker, not property(Subject, Factive, Presupposition)))
and factive(Factive)
=>
                bel(Speaker, bel(Subject,Presupposition)).
```

A.3 Implicature

```
% follows from Quantity:
%  'the utterance of such a complex sentence implicates that
%   both the constituent sentence and its negation are compatible
%   with what the speaker knows.' [GAZDAR79, p61].
default fc(1,S,U1,U2) :
            utt(S, or(U1,U2))
            =>
            imp(S,U1).
default fc(2,S,U1,U2) :
            utt(S, or(U1,U2))
            =>
            imp(S,U2).
default fc(3,S,U1,U2) :
            utt(S, or(U1,U2))
            =>
            imp(S,not U1).
default fc(4,S,U1,U2) :
            utt(S, or(U1,U2))
```

```
            =>
            imp(S,not U2).

% implicatures are believed by default:
default sensible(S,U) :
            imp(S,U)
            =>
            bel(S,U).
```

A.4 Rationality

```
% We are informed of knowledge that is mutually known in
%  the community:
default aware(Agent,A) :
      mutual(A) => bel(S,A).

% If we believe the antecedent of a rule, we believe its consequent:
default implicit(Agent,A,C) :
      mutual(=>(A,C)) and bel(Agent,A) => bel(Agent,C).

% If we believe a list, we believe the items:
default conjunct(Agent,List,X) :
      bel(Agent,List) and member(X,List) => bel(Agent,X).

% Positive Introspection (patently left-recursive):
default pos_int(Agent,B) :
      bel(Agent,B) => bel(Agent,bel(Agent,B)).

% Motherhood...
fact member(X,[X|Tail]).
fact member(X,Tail) => member(X,[H|Tail]).

fact mutual(=>(A,not property(O,B))) and bel(S,A)
      => not bel(S,property(O,B)).

% Re-write rules:
fact bel(S,property(O,not B)) => bel(S,not property(O,B)).
fact bel(S,not property(O,B)) => bel(S,property(O,not B)).

fact imp(S,not property(O,B)) => imp(S,property(O,not B)).
```

A.5 Miscellaneous

A.5.1 World Information

```
% WORLD INFORMATION:
% definition of bachelor:
fact mutual(=>(property(X, bachelor),
        [property(X, male),
            property(X, adult),
            property(X, not married)])).

% definition of spinster:
fact mutual(=>(property(X, spinster),
        [property(X, female),
            property(X, adult),
            property(X, not married)])).

fact mutual(=>(property(Anyone, female), not property(Anyone, male))).
fact mutual(=>(property(Anyone, bachelor), not property(Anyone, spinster))).
```

A.5.2 Lexical Information

```
% The non_criterial facts:
fact nonc(bachelor,male).
fact nonc(bachelor,adult).

fact nonc(spinster,female).
fact nonc(spinster,adult).

% The factive facts:
fact factive(regret).
```