Abstract

Title of Dissertation: A View of One's Past and Other Aspects of Reasoned Change in Belief

Michael J. Miller, Doctor of Philosophy, 1993

Dissertation directed by: Associate Professor Donald Perlis
Department of Computer Science

This dissertation reports work on *reasoned change in belief*, specifically related to the following concepts:

- pronominal indexicality in first-order logic (FOL)
- typicality and range defaults
- terminological (language) change over time

Results are presented in each:

(i) The pronominal indexical "I" – its meaning changing with who is speaking – is given formal treatment in the context of a logic puzzle, solving a problem previously posed in the literature.

(ii) A new form of default information in which typicality is viewed as spreading over a *range* of possible default conclusions is isolated. "Cardinals are typically red or russet" is a reliable default while both "cardinals are typically red" and "cardinals are typically russet" are not. The range "red or russet" is essential, though shown to require adjustment of previous formalisms.

(iii) Terminological change over time, the process of language flexing on the fly as new terms and new meanings become important to a reasoner, is formalized in the context of mistaken past beliefs. This process often is spurred on by *contradictory beliefs*, which are viewed here as positive aids to reasoned change; a result is proven on *recovery* from contradictions as well.

We concentrate on the latter theme, specifically change in meaning and language usage over time (chapters 1–5); the indexical and default results are presented separately in chapters 6 and 7, respectively. The main technical contributions are in chapters 3–7. In 3 we introduce new concepts for terminological change. In 4, a general theorem about step-wise reasoning in time when contradictions are present is proven. In 5, a step-wise formalism that can handle specific problems of terminological change is presented. In 6, a first-order logic treatment of the first person indexical "I" is given. In 7, some apparently new difficulties in reasoning about typicality are uncovered.

A View of One's Past and Other Aspects of Reasoned Change in Belief

by

Michael J. Miller

Dissertation submitted to the Faculty of the Graduate School
of The University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1993

Advisory Committee:

Associate Professor Donald Perlis, Chairman/Advisor
Assistant Professor Bonnie Dorr
Professor Jack Minker
Associate Professor James Reggia
Professor Ellin Scholnick

Dedication

To the memory of my father.

Acknowledgements

There are many people to whom I owe a great deal of gratitude for their support, advice, help, and encouragement during the course of work on this thesis. My deepest gratitude is to my thesis advisor, Dr. Donald Perlis. Long ago Don introduced me to logic and he has inspired me ever since. His teachings are always clear; his criticisms wise and insightful and always welcome. Most of all his understanding, encouragement, (gentle) prodding, and faith in me and my work kept me going long past when I was about to give up. Remarkably his patience never seemed to dwindle, though I gave him many an opportunity!

I would also like to thank the members of my thesis committee – Professors James Reggia, Jack Minker, Bonnie Dorr, and Bill Gasarch of the Department of Computer Science and Professor Ellin Scholnick of the Department of Psychology – for their time, effort, and criticisms. Dr. Minker's comments on an earlier draft of this thesis were unusually detailed and careful, and his insightful questions have since provoked further research and results.

Numerous fellow graduate students and departmental staff have contributed to this work with their support, both technical and personal. Among them are Parke Godfrey, José Alberto Fernández, Yuan Liu, Lynne D'Autrechy, Granger Sutton, and Malle Tagamets. Special thanks to Nancy Lindley; her job is to keep the graduate students on track – an unenviable task!

I am grateful to the National Science Foundation (grants OIR-8500108 and IRI-9109755), the U.S. Army Research Office (grants DAAG29-85-K-0177 and DAAL03-88-K0087) , and the Martin Marietta Corporation for their support of my research.

I am unable to express how much support I have received from my friends, and even more from my family. One's family and close friends seem to bear the bulk of any burden cast by such an endeavor; mine certainly did. Yet they stood beside me throughout, never once shying away, always offering comfort. Thank you.

Finally, I would have never written this thesis had it not been for Lori Singer. My interest in Computer Science was piqued because of her and my desire to pursue that interest was encouraged by her. Beyond that, she taught me much over the years for which I remain indebted, though of that she is likely unaware.

M.J.M.

Table of Contents

List of Figures

A View of One's Past and Other Aspects
of Reasoned Change in Belief

Michael J. Miller

April 21, 1993

Survey of earned doctorates???
CHECK BIB AND EPI. PAGE no. in TOC
CHeck ALL page nos. And FIG nos.
Complete acknowledgement section
Where do grant acknowledgements go? (VS grant)
run SPELL CHECK

This comment page is not part of the dissertation.

Typeset by LaTeX using the `dissertation` style by Pablo A. Straub, University of Maryland.

Prologue (Aims and Accomplishments)

This dissertation reports work on *reasoned change in belief*, specifically related to the following concepts:

- **pronominal indexicality in first-order logic (FOL)**
- **typicality and range defaults**
- **terminological (language) change over time**

Accomplishments include the following: (i) The pronominal indexical "I" – its meaning changing with who is speaking – is formalized in the context of one of the *Knights and Knaves* problems [Smullyan, 1978], solving a problem previously posed in the literature. (ii) A new form of default information is isolated in which typicality is viewed as spreading over a *range* of possible default conclusions. "Cardinals are typically red or russet" is a reliable default while both "cardinals are typically red" and "cardinals are typically russet" are not: the range "red or russet" is essential. Proper representation of such "range defaults" is shown to require adjustment of previous formalisms. (iii) Terminological change over time, the process of language flexing on the fly as new terms and new meanings become important to a reasoner, is formalized in the context of mistaken past beliefs. This process often is spurred on by *contradictory beliefs*, which are viewed here as positive aids to reasoned change. A variety of novel knowledge-representation tools are developed for this treatment. A theorem is proven giving sufficient conditions for *recovery* from contradictions within the formalism. The formalism is applied to problems in commonsense reasoning and is shown to give intuitively plausible results.

The research reported in this dissertation can be viewed as a series of inroads toward achieving the long range goal of building an integrated reasoning system which has abilities suited to an advice-taker, such as Tommy in this dialogue:

(1) TOMMY: "Look Mom, Jane's feeding the birds!"

(2) MOTHER: "That's not Jane. It's her twin sister."

(3) TOMMY: "Oh, her twin sister."

(4) MOTHER: "Do you know what kind of birds they are?"

(5) TOMMY: "No."

(6) MOTHER: "They're cardinals."

(7) TOMMY: "But they're not red. I thought cardinals were bright red like my bike."

(8) MOTHER: "I once thought that too. But only adult males are red. Females are russet."

(9) TOMMY: "So they're really cardinals, even though they're not the color of my bike."

(10) MOTHER: "That's right."

(11) TOMMY: "How come only females are here?"

(12) MOTHER: "I don't know. That's odd isn't it?"

(13) TOMMY: "Maybe Jane's sister knows. Let's ask when she's done feeding the birds."

The natural-language aspect here is not our focus. Rather it is the underlying change of beliefs implicit in the reasoning done by Tommy.

The key features of the dialogue which have attracted our attention are: (i) the use of the (first person pronominal) indexical "I" in lines 7, 8, and 12; (ii) the use of a newly identified kind of default, a *range default* (described shortly) in lines 8, 11, and 12; and (iii) the assessment and correction of mistaken past beliefs, especially when the mistakes are signaled by conflicting beliefs, when the mistakes involve object-identification errors (e.g., Tommy's misidentification of Jane's sister in line 1), and when the correction is accompanied by *terminological change* (e.g., Tommy's belief about Jane, reported in line 1, is later in line 13, seen by Tommy as being more appropriately about Jane's sister.)

The following summarizes our results and is exemplified by references to the above Tommy scenario.

• Pronominal Indexicality

An indexical is an expression whose referent is dependent on the context in which that expression is used. "I", "now", and "here" are examples. Few would argue that indexicals are

not prominent in commonsense reasoning, yet there has been almost no work on them in that domain. Most of the work has been in the area of natural language processing (NLP) instead. While some of our motivating examples have an NLP flavor, this is not necessary as other examples show.

Our first effort represented a head-on confrontation with issues in knowledge representation and indexicals varying over context of speaker. In reasoning about utterances by multiple agents, it is essential to be able to interpret the meaning of the utterance with respect to the utterer. Two utterers may use the same word to mean different things. This is particularly obvious in the case of the pronoun "I".[1] We designed a series of axioms which are appropriate to the reasoning Tommy (and Mother) would carry out to properly interpret the indexical (changing) meaning of "I" (lines 7, 8, and 12) in the scenario above and employed them in the representation of, and solution to, one of Smullyan's logic puzzles.

This puzzle had already been treated in first-order logic (FOL), by Ohlbach [Ohlbach, 1984], but for him proved very tricky and, as he noted, his formalism had a highly unintuitive flavor. We show that the confusion was due to a failure to take account of the proper role of indexicality. Specifically, when an agent utters an expression containing "I", such as "I am a knight", the meaning (truth-value) of this depends on who the agent is. Ohlbach's formalization did not fully account for this in a clear way. To get such information into the FOL framework is made more complicated by the problem-definition, which requires knights to be able to tell *only* true statements. Several new predicates were required in order to properly represent the problem. We eventually arrived at a resolution-proof of the desired solution to the problem that was far more intuitive than Ohlbach's. The result was a combination of techniques from theorem-proving and natural-language processing. This work was reported in

[1] We confer no special status here upon 'I' over other indexicals like "you", "now", "here", "this", "that", etc (but see [Frege, 1956], [Perry, 1977], and [Perry, 1979]). AI must address them all. We were just following Ohlbach in treating "I". Step-logics (see below) begin to address "now", and the work reported here in chapters 3 and 5, and in [Miller and Perlis, 1993a] touch on aspects of "this" and "that" vis-à-vis the seemingly indexical nature of proper names. See [Hirst, 1981] for a survey of literature on determining reference in indexical contexts.

[Miller and Perlis, 1987b] and [Miller and Perlis, 1987a], and appears here as chapter 6.

• Defaults

It is widely accepted that default (or non-monotonic) reasoning is endemic in commonsense reasoning. Defaults can be, and often are, viewed as typicality statements of the form "P's are typically Q's". A very different way to view typicality (and hence defaults), and the intuition behind the treatment examined here, is to treat a reasoner's mental concept of a *typical* or *generic* instance, which roughly corresponds to a general (indefinite) description, as an object in its own right. For example, I may have a "mental notion" of what is for me a typical tree. That this typical tree notion (for me) has "leaves" encodes my default that "trees typically have leaves"; that it has "branches" encodes my default that "trees typically have branches".

We attempted to formalize this intuition by extending a first-order language to include representations of these mental notions, in the form of constant symbols called *typicality-* (or simply, *typ-*) constants, which are written as typ_Φ for expressions Φ in the language associated with an indefinite description. As reified objects of thought *typ*-constants have properties (this is how we encode defaults) and are subject to manipulation in the reasoning process. We prove two simple theorems showing that *typ*-constant default reasoning is transitive and composes with logical implication. These properties show that *typ*-constants lead easily and intuitively to certain desired default conclusions.

Based on examples with *typ*-constants, we then noted numerous cases in which defaults must be specified as a *maximal range that cannot be reduced*. These so-called "range defaults" are accepted defaults of the form "P's are typically Q's" where Q is a disjunction and for every shorter disjunction S formed from the (disjunctive) components of Q, the (sub-)default "P's are typically S's" is rejected. Range defaults are important in commonsense reasoning whenever a category consists of more than one major subcategory. For instance, cardinals are almost all either red or russet in color, with large numbers of each, and exceptions being rare. So the default "cardinals are typically red or russet" cannot be reduced to a narrower default. That is both "cardinals are red" and "cardinals are russet" should be rejected as defaults. This is important for commonsense reasoning. To note that a gathering of only russet-colored cardinals seems odd (line 12 in the Tommy scenario above) Mom must be aware of the range of

typical cardinal coloring (red or russet) and that the range cannot be narrowed. Moreover, to correct an overly restricted default ("cardinals are red") to a more appropriate range default ("cardinals are red or russet") requires a mechanism for *asserting the inappropriateness* of the former default.

Properly representing range defaults and the inappropriateness of restrictions is not completely straightforward. Although we first observed this for *typ*-constants, the phenomenon is quite general to all default formalisms, and yet is unexplored in the literature. We prove, for instance, that it cannot be done directly in Reiter's Default Logic. Circumscription (using *ab*-predicates) also has difficulties with ranges: we prove that *negating* an inappropriate default produces merely a counter-example assertion and does not express that the default itself is inappropriate. We prove range defaults expressed via *typ*-constants lead to actual contradictions. We present a formal proposal to solve the range default problem and discuss its shortcomings (which we have since successfully overcome in [Miller and Perlis, 1993b].) This research was reported in [Miller and Perlis, 1991], and appears here as chapter 7.

• Terminological (Language) Change

The tie between linguistic entities (e.g., words) and their meanings (e.g., objects in the world) is one that a reasoning agent had better know about and be able to alter when occasion demands. This has a number of important commonsense uses. The formal point is that a new treatment is called for so that rational behavior via a logic can measure up to the constraint that it be able to change usage, employ new words, change meanings of old words, and so on, over time.

The usual fixed language with a fixed semantics that is the stock-in-trade of AI seems inappropriate to this task. Here we propose "active logics" based on the step-logics of [Elgot-Drapkin and Perlis, 1990]. A step-logic models belief reasoning by describing and producing inference one-step-at-a-time, where the time of reasoning is integral to the logic. Complicated reasoning made of many successive inferences in sequence take as many steps as that sequence contains. Step-logics are inherently non-monotonic: theorems can disappear at every inference step. Indeed, in the version most-investigated to date, SL_7, some theorems do always disappear at every inference. These "necessarily-disappearing" theorems are the time-stamps: wffs of

the form $Now(i)$ where i is a step-number giving the current time. $Now(i)$ is updated to $Now(i+1)$ as inference proceeds from time i to time $i+1$.

We have applied active logics to the specific issue of terminological change vis-à-vis mistaken beliefs. E.g., Tommy's belief about Jane feeding the birds (line 1, above) is later seen by Tommy as being more appropriately about Jane's sister. Tommy has committed an object-identification error of the so-called "compression" type wherein a singly denoting term (e.g. "Jane") is inadvertently used to refer to more than one object (i.e., Tommy's initial use of "Jane" can be viewed as referring to both Jane and her sister).

The change in belief occurs here after his mother (Tommy's advice-giver, or *tutor*) isolates his mistake and informs him of its nature by distinguishing *that* (person seen) from Jane (line 2). We develop a formalism which has expressions for such natural language demonstratives as "that" in the Tommy scenario. We call these formal expressions *reality terms*; their function is to denote an object which was "presented" to the agent and (possibly) incorrectly identified by her in the past. Thus these are terms used to denote an entity, (possibly) replacing a previously held but incorrect description of that same entity. In the example scenario, the word "it" (line 2, second sentence) also serves the role of a reality term, and the object referred to both by *that* and *it* in line 2 is subsequently identified with Jane's sister.

By informing Tommy of the distinction between *that* (person seen) and Jane, Tommy's mother supplies him with a partial means to correct his own past mistaken beliefs. He uses her advice to revise his beliefs accordingly, correctly coming to believe that Jane's sister is feeding the birds (line 13). We develop formal tools for reality terms useful for expressing object-identification errors like Tommy's, and we incorporate into step-logics a mechanism for correcting such errors once tutorial advice, like his mother's, is given.

Contradiction and conflict play a key mediating role in the reasoning here, serving to signal (to Tommy) that (his) past beliefs must be re-assessed and revised (lines 1-2 and 7). In most formal AI, contradictions are anathema since most logics become useless in their presence. However human reasoning is not usually thrown into such disarray by contradictions. Thus we have sought formal ways to be more accommodating of contradictions. Little more than lip-service has been paid to the treatment of contradictory information in commonsense reasoning. Probably this is due to the customary reliance on standard logics having the "ex contradictione

quodlibet" feature: from a contradiction all is entailed. In Elgot-Drapkin's work, this is called the "swamping" problem. There are non-standard logics, the paraconsistent logics, that do allow contradiction without swamping; however, in commonsense reasoning one wants not only to avoid swamping but also to somehow undo or at least cease believing the contradiction. Earlier step-logic work had a way to *ignore* contradictions. But more is needed. Not only must we adjudicate between contradictands, we must also prevent earlier mistaken beliefs (revealed by contradiction) from infecting *future* reasoning. Conflicting beliefs, mistaken beliefs, and their consequences must be controlled, so as not to infect other beliefs indefinitely into the future. Note that Tommy *recovers* from his conflicting beliefs when the conflicts (about Jane and about cardinals) no longer adversely affects his beliefs (presumably, by line 13); not only does he accept Mom's advice (lines 2 and 8), but he also rejects a *consequence* of his earlier mistaken view (see lines 7 and 9, concerning his bicycle).

Recovering from contradiction was broached in Elgot-Drapkin's work, but only in an ad hoc way. There a conjecture was formulated, to the effect that, under (unspecified) circumstances, a step-logic should be able to regain consistency from an initially inconsistent set of beliefs. Here (in chapter 4) we begin to make inroads, in a limited way. We develop new step-logics which under suitable conditions are shown to recover from *direct contradictions* and their consequences (our *dc-recovery* theorem). This amounts to importing much of a truth-maintenance, or belief revision, system *into* the logic, which then – unlike a usual belief revision system – operates *during and as part of* the ordinary reasoning of the logic. This means that world knowledge can be brought to bear on the truth-maintenance (belief update) process, and other reasoning need not be halted while the belief updating is occurring. We advance two postulates concerning commonsense reasoning, the *short-chain* and *lazy-corroboration* hypotheses, which keep in check the computational bookkeeping required by our dc-recoverable step-logics.

These additions to step-logics, namely the mechanisms enabling terminological change and those for *dc*-recovery, have allowed us to solve commonsense problems centered around object-identification error. Two such problems are solved in detail here; one we call the *Mistaken Car* and the other *Two Johns*. The former problem involves an identification error much like Tommy's misidentification of Jane. The latter problem, *Two Johns*, compounds this with the introduction of an ambiguity: both the incorrectly identified object and the object it was

(incorrectly) *taken to be* share the same name until later disambiguated with reality terms.

Thus our formalism allows contradictions and *benefits* from them to stimulate belief re-assessment; we have proven (in suitable settings) recovery from contradictions; and we have treated object-identification errors. All these can occur together in a mutually beneficial way to produce intuitive commonsense behavior in our formalism.

To our knowledge, treatment of contradictions with recovery over time, of *reality terms*, and of object-identification errors (of compression have not been developed elsewhere, though they have been raised as important topics in the literature ([Maida, 1991], [McCarthy and Lifschitz, 1987], and [Harman, 1986]). This work is reported in [Miller and Perlis, 1993a], and has been the focus of the bulk of the research reported in this dissertation. It appears here as chapters 1–5.

The main technical contributions are in chapters 3–7. In chapter 3 we introduce our formal treatment of reality terms and the tutorials which assert object-identification errors. In 4, the *dc*-recovery theorem about step-wise reasoning in time when direct contradictions are present, is proven. In 5, a step-wise formalism that can handle problems such as the *Two Johns* problem above is presented.

As mentioned, the indexicality and default works are presented, as separate endeavors, in chapters 6 and 7, respectively.

The early chapters (1–3) serve the dual purpose of reviewing the various literatures relating to these themes and setting the stage for our formal treatment of *Two Johns*-like problems. The progression will be from informal and intuitive to formal. Chapter 1 is motivational leading up to the discussion of belief and belief reasoning, discussed in chapter 2. Chapter 3 focuses on error, as viewed here, closing with a semi-formal presentation of *Two Johns*.

• Implementation Work

The following implementations related to the research reported here were carried out in Prolog:

1. The indexicality axioms for the pronoun "I" were implemented in a context-free parser.

2. The terminological change work was implemented as an addition to step-logics, and used to solve the *Mistaken Car* problem.

3. A step-logic with the dc-recovery property was implemented.

4. A step-logic "decay" mechanism which addresses computational space concerns was implemented.

Chapter 1

General Background

1.1 Introduction

The same premise underlies this dissertation as has motivated much of the large artificial intelligence (AI) research effort investigating commonsense default reasoning formalisms. That premise is: the commonsense world is far too complex for reasoners, human or otherwise, to be aware of all facts and information that may be salient to a situation at any given time. As noted in [Etherington, 1988], a consequence of this premise is the so called "qualification problem" [McCarthy, 1980]:

> Virtually none of the decisions one makes everyday are made with complete certainty. With little effort, an endless supply of more or less probable scenarios can be constructed that contraindicate any chosen course. Yet people are not paralyzed by indecision; they continue to act and to decide in spite of all this uncertainty. [Etherington, 1988]

Researchers have addressed these concerns by constructing default reasoning formalisms that somehow render meaningful a reasoner's ignorance or lack of relevant knowledge (e.g., see [Reiter, 1978] [McCarthy, 1980], and [Reiter, 1980]). In brief, these formalisms offer various solutions to the problem of representing how a reasoner might, when necessary, jump to a reasonable yet defeasible conclusion based on whatever knowledge is available plus some default rule(s). Just what *is* a reasonable conclusion, how it is reached, and what the default rules are varies from formalism to formalism and is an issue that need not overly concern us here.

Instead we shall focus on an inevitable consequence of reasoning with incomplete information or by default, namely mistaken or erroneous beliefs.

Error is intimately related to one feature common to all default reasoning formalisms: *non-monotonicity*. Intuitively, non-monotonic reasoning amounts to this: a reasoner with less knowledge or information might draw conclusions which that same reasoner with more information might not draw. That is, the acquisition of information might invalidate the application of a default rule. In the case of a human reasoner this is especially apparent when we consider reasoning as a process occurring in real time. We often draw a conclusion, say α, based on a set of beliefs, Γ, and later come to believe $\Gamma \cup \{\beta\}$ from which $\neg\alpha$ follows. When this happens, we *sometimes* note our mistake (e.g., that α was believed in the first place and why was it wrong to do so) and take some appropriate corrective action (e.g., stop believing α and the consequences of $\Gamma \cup \{\alpha\}$). Here α is an example of a former mistaken belief, as the term is intended in this work.

Though it is most familiar to the AI community, default reasoning is not the only cognitive process that gives rise to error and mistaken beliefs. Perception-based beliefs, those constructed directly from one's percepts (e.g., the percept of a particular color) without the apparent intervention of any conscious reasoning process, may be misleading. This is common in the event of illusion, unusual lighting conditions, poor acoustics, and the like. (See the discussion of the appearance-reality distinction in section 1.5.1.) Other beliefs, those formed in accord with information supplied by others, may be erroneous when the source is deceitful, misinformed, or has in mind a different meaning (for a word or concept, etc.) than the receiver of the information understands. Reasoning-based beliefs may be erroneous or mistaken when the line of reasoning is fallacious or (sound) reasoning is reliant upon other mistaken beliefs. Still other false or mistaken beliefs arise because of faulty recall of past facts and events or subconscious affirmation of facts and events occurring only in dreams, and so on. In short there appear to be a multitude of ways that the human cognitive system is subject to the occasion of erroneous beliefs. Whether the reason for this susceptibility is an evolved efficient means of jumping to reasonable, usually trustworthy, beliefs that aid in survival, as might be the case in default reasoning, or a slight hardwiring mixup or faulty storage-retrieval process in the brain, as might be the case when, on the odd occasion, dream events are taken to have occurred during

the dreamer's conscious state, is not so much at issue here. Rather we are concerned that the human cognitive system, the only known example of an intelligent reasoner, *is* subject to error. Moreover, the inaccuracies and frailities of that system, the multitude of sources of misinformation, incomplete information, deception and illusion that any intelligent system is likely to confront all point to a seemingly inescapable consequence: intelligent artificial agents will be forced to confront mistaken beliefs.

The AI community has already seen the need to build belief revision modules into "intelligent" programs as a fix for default-reasoning-based-errors. For instance, the, truth maintenance system (TMS) of [Doyle, 1979] and its assumption-based counterpart (ATMS) of [deKleer, 1986] are systems tuned to revise and manage belief sets in accord with an artificial reasoner's temporally changing base of available information, perhaps somewhat like the cognitive belief maintenance that people continually need to perform in response to noted erroneous beliefs. A traditional system such as TMS is a program which works together with another, say theorem proving, program. The job of the belief revision module might be to keep track of beliefs that are currently labeled "IN"(presumably, believed) and those that are currently labeled "OUT"(presumably, not believed) – perhaps regardless of whether they had been previously held – and revise the sets "IN" and "OUT" as necessary, in accord with newly acquired information as supplied by the theorem prover or (simulated) observation module of the modeled reasoner. Significant as this traditional AI version of the belief revision process is to commonsense reasoning, it neglects any kind of reasoning *about* mistaken beliefs, which turns out to be a critical aspect of commonsense reasoning. The ultimate fate of the mistaken belief is retraction from the current belief set, "IN". What of the knowledge that the belief was once held? That information may not be available to the modeled reasoner. And what of the knowledge representing what was wrong with the erroneous belief in the first place? That, too, is information not available in traditional systems. In short, they have no conception or view of their past, and in particular, of their past errors.

On the other hand, humans are often aware that their apparently false and questionable former beliefs were once thought to be true, and often have knowledge about what was wrong in believing them in the first place; all this is in addition to correcting their errors (i.e., traditional belief revision). Error itself is a topic that we are able to reason about.

Cognitive psychologists believe that by the age of 3 to 5 years children acquire the ability, or have sufficiently refined the cognitive mechanisms, to do just this sort of reasoning. Young children develop the facility to compute mental belief states based on simultaneous representations of both a former (or fictional, etc.) view of the world, represented as once believed correct (or pretended, in the event of pretense, etc.) and which is currently believed to be mistaken (or fictional , etc.), and a currently accepted view of the world, represented as the accepted actual state of the reasoner's world [Astington and Gopnik, 1988]. This allows children (and adults) to perform some metacognitive tasks that are very basic to their intelligent understanding of the world and their role in it, thereby making possible the very useful ability to explain behaviors that had been (mis-)guided by apparently mistaken beliefs, including the behavior of accepting other beliefs as true. (A brief discussion of these metacognitive tasks, in which examples are given, appears in section 1.5. For more on this see [Astington *et al.*, 1988].)

In broad terms the position advanced in this thesis is that intelligent robots, and other formal reasoning systems, will inevitably be confronted with erroneous beliefs, and therefore the truly intelligent robot will be equipped to consider the possibility of error. *Moreover it will be able to represent and reason about its own past beliefs*, whether those beliefs are taken to be mistaken or not. It will question some truly false beliefs and not question others. It will even question and doubt some true beliefs, much as we do. All of this is necessary if it is to have an intelligent understanding of the world and its role in it.[1] In particular, a truly intelligent robot must detect and represent certain of its own past beliefs as being apparently false or questionable, recover from its mistakes by revising its beliefs, and, at times, come to understand why a mistaken belief was held in the first place.

A first step toward endowing artificial reasoners with this facility, and the one advanced here, is a *formal simultaneous representation of two views of the world*; a past mistaken view, and a current accepted view.

[1] This in addition to the robot's need to reason about *other's* correct and mistaken beliefs.

1.2 A Note on Terminology

Before proceeding it will be helpful to offer an intuitive feel for some terminology that will be used throughout. This will all be made more precise shortly but this introduction will serve to allay some confusion. The following terms will repeatedly be used: current (or present), past, former, and mistaken (or erroneous) beliefs.

Current beliefs are regarded as those beliefs which are held "now"; *past* beliefs are those beliefs which an agent once held (and possibly still holds "now"); *former* beliefs are past beliefs which are not currently held. A *mistaken* (or *erroneous*) belief is a past (and possibly former) belief which the agent currently believes was produced in virtue of some past mistake, such as a perception-based identification error. For an agent to believe one of her former beliefs was mistaken is a psychological stance independent of the actual truth or falsity of the belief in question. Like any other belief, the judgement of truth or falsity here is seated squarely on the shoulders of the agent of interest, not on an outside party who is privy to more or less information about the world. There is no concern here with an omniscient oracle's verdict on the veracity of a reasoner's beliefs. Rather, the concern is with the psychological issue of what a reasoner *takes* to be real or correct as opposed to false, illusory, etc., and not what is *actually* true or false.[2] Note also that "true" or "false" need not be permanently assigned to one's beliefs. To the contrary, we change our minds about the apparent correctness of our beliefs continually as we learn about the world.

From time to time we will also refer to an agent's view of her own past reasoning. What comprises this view? Beliefs: current beliefs about past (and former) beliefs.

Beliefs of all sorts can be regarded as propositional attitudes which bear heavily on agent action and intentions. When one believes that α, she generally will act in accordance with that belief. If I believe I see a wolf up ahead, then if possible I'll avoid it. If I see my friend deliberately walking toward the wolf, then her actions may cause me to think that she isn't afraid of wolves, or she doesn't see it, or maybe she thinks it's a dog, not a wolf; and so on. My belief that the critter is a wolf becomes a past belief as time passes. When my friend convinces me that the animal up ahead is actually a dog, my thought that it is a wolf becomes a former belief. Now, because I view my past belief as mistaken I, too, no longer fear the animal.

[2] See [Barwise and Perry, 1983] for a discussion of the related issues of *cognitive* and *external* coherence.

The description above of a view of one's past is mostly open-ended. A view may include, at least: (i) information about which beliefs were once held and, perhaps, how they came to be held; (ii) information about certain beliefs which were not held; and specifically that they weren't held at a given time (i.e., time-situated negative introspection); and (iii) information about what later seem to be mistaken. The concern in this dissertation is mostly directed at the last of these components, though not to the exclusion of the others. The sort of view we have in mind contains information useful for describing, explaining, reasoning about, understanding, and so on, the mistakes that one sees reflected in their past beliefs.

One component that such a view may contain is simply that a given belief was once, but is no longer, held. A pseudo-formal paraphrase of the above wolf story is:

$$Bel(I, \exists x Wolf(x) \wedge UpAhead(x), before(now)) \ \wedge \ \neg Bel(I, \exists x Wolf(x) \wedge UpAhead(x), now)$$

i.e., "I previously (but no longer) believed there to be a wolf up ahead." Even though information like this is useful to a reasoner, it alone, is insufficient for many purposes. We shall see in chapter 3 what more is needed.

A caveat: we will shortly discuss "false beliefs" as that expression is sometimes used in cognitive psychology. That use differs from the above notion of mistaken or erroneous beliefs.

1.3 A Motivating Example

Beliefs that are later judged to be mistaken are quite prevalent in commonsense reasoning. It shouldn't take much to convince one of that. The story about the wolf in the previous section might be enough to do the convincing, but of what significance is the use of having a view of past mistakes? Consider another motivating example which should be all too familiar to nearly everyone who has ever owned an automobile.

> The Mistaken Car: Most of us have had the disconcerting experience of misidentifying our car in a parking lot. You approach a car in the lot thinking that it is your own, but when you try to unlock the car's door you fail. After convincing yourself that you are using the correct key you might notice an unexpected dent in the car's fender or some unfamiliar personal belongings inside the car which leads you to suspect that the car is not your own, but a look-alike that, perhaps, you hadn't

even known existed before now.[3] Once you have enough evidence to support this growing suspicion you come to believe that, indeed, you had mistaken another's car for your own.

We will not delve too deeply into this example for the moment. Rather its purpose here is to enable a discussion of some of the implications that might result if we were unable to reason about our past beliefs. The consequences would be more than an inconvenience or annoyance in the commonsense world; they can be disastrous – consider that the misidentified car's owner may be prepared to use a gun if she does not receive a satisfactory explanation of what *appears* to her to be your thievious behavior. Providing such an explanation requires reasoning about your past belief; that you thought *this* was your car. Having no such reasoning capability means having no explanation.

In addition to being unable to explain many of their behaviors, reasoners without the necessary mechanisms to recall their past errors might be mistaken belief recidivists: they might repeat the same mistake over and over. (Or perhaps they would come to rely on some kind of "unconscious" learning.) "Conscious" (or explicit/declarative) learning from past mistakes seems to presuppose that one take note of and recall those mistakes.

One counterpart of repeated mistakes might be that commonsense reasoners would find it difficult to restructure their defaults of typicality about the commonsense world. Consider the course of development of a child's defaults about birds: After repeatedly hearing her parents say "Look, there's a cardinal!", while pointing to a mature male cardinal, the child may come to believe that all cardinals have bright red feathers. The child, though now able to correctly identify some cardinals, will likely misidentify the female and immature of the species. Upon seeing some female cardinals, misidentifying them, and being told that she has misidentified them, she ought to begin to suspect the validity of her belief regarding cardinal color and come to believe the default "cardinals are *typically* red". Indeed, upon seeing enough female cardinals she ought to replace her initial belief with the more accurate default "cardinals are

[3] That is, you hadn't known that *this* particular car existed even though you might well have believed that look-alikes for your car exist. The idea here is that upon the encounter with the misidentified car you had no identifying expression or name for the car, or more precisely that you are ignorant of having a name for the car. This assumption is made only to make the example more robust in later discussions.

typically red or russet".[4] A reasoned change in belief like this can occur by virtue of the sheer number of misidentified cardinals, it seems, only if the child can note that her initial belief is erroneous.

In short, there seems to be a basic epistemological advantage gained from one's ability to reason about her mistaken beliefs: it helps her to better know her world. Flavell, Green and Flavell [Flavell *et al.*, 1986] in their work on the development of the *appearance-reality distinction* mechanism in children, a metacognitive stance akin to reasoning about mistaken beliefs, say:

> The distinction arises in a very large number and variety ecologically significant cognitive situations. In many of these situations, the information available to us is insufficient or misleading, causing us to accept an apparent state of affairs (appearance) that differs from the true state of affairs (reality). We are variously misled or deceived by the information receive from or concerning people, objects, actions, events, and experiences. (p.95) ... all systematic pursuit of knowledge presupposes at least some awareness of the appearance-reality distinction. ... Although we may not know that appearances have in fact deceived us in any specific cognitive situation, we do know as a general fact that such deception is always possible. That is, ... we have acquired the metacognitive knowledge that appearance-reality differences are always among life's possibilities. (p.96) [Flavell *et al.*, 1986]

Reasoned change in belief then seems essential.

1.4 Goals

In addition to giving a preliminary understanding of the sort of view of one's past that we have in mind here, the *Mistaken Car* story should also suggest several provocative questions, including:

---

[4] Even if she has never been told explicitly that the female cardinals have russet feathers and the male cardinals have red feathers. We call defaults of the form "P's are typically Q's" where Q is a disjunction and for every shorter disjunction S formed from the (disjunctive) components of Q, the (sub-)default "P's are typically S's" is rejected *range defaults*. Range defaults are discussed in chapter 7.

(1) How does one come to believe that α in the first place?

(2) How does one come to suspect, and then decide, that a belief is mistaken?

(3) How does one determine exactly *what* error has been made?

(4) How does the situation unfold, representationally, in one's head?

(5) How does one's belief set change during the unfolding of the situation?

  (a) Which beliefs are modified and how?

  (b) Which beliefs are retracted?

The primary aim chapters 1–5 of this dissertation is to begin to develop answers to these, and related, questions. This means we must come to understand precisely what are the processes of coming to have a view of one's past and of reasoning about former beliefs, what are the cognitive requirements for them, and what faculties the ability to have such a view and reason about error confers upon reasoners. To this end we start by drawing together research from various disciplines that touch on the general topics of belief and error. Once we have (some of) the answers to the above questions we can begin to develop formal computational tools and techniques for representing one's past and reasoning about former beliefs. We will see that the basic formalism must provide a reasoning agent with (1) a dynamic internal model of the agent's evolving beliefs about the world, and (2) a flexible formal language with which the agent can reason about that model in relation to her current view of the world.

1.5 Cognitive Psychology

Perhaps the largest body of extant literature concerned with an agent's view of her own (and other's) beliefs comes from developmental and cognitive psychology. Much of the relevant research reported in that literature is an outgrowth of the seminal work done by Flavell, Green, and Flavell on the development, in young children, of the metacognitive ability to make the so-called *appearance-reality distinction* (ARD)[Flavell *et al.*, 1986]. The outgrowth of literature is aimed at understanding the acquisition and development, in young children, of the ARD and at least three other, seemingly related, metacognitive tasks or abilities. (See [Astington *et al.*, 1988] for a collection of some of the recent literature on these metacognitive tasks.) The apparent relationship between at least three of these four tasks manifests itself

most notably in their near simultaneous development in children by age 3 to 5 years.

The facility humans have with these tasks is a motivational force for the current work. One of the tasks, *representational change*, is in evidence in the *Mistaken Car* story. Let us briefly look at these tasks.

1.5.1 Four Metacognitive Tasks

Representational change refers to one's ability "to understand that one's representation of an object or a phenomenon has changed, and to remember the previous representation"[Astington and Gopnik, 1988] (p. 193). A paraphrase of a representational change (belief) report, as given by Astington and Gopnik (p.193), is: "I used to think x but now I know y".[5] Thus a report of representational change is a report stating that one's own former belief is, at the time of the report, taken to have been mistaken, at the time the belief was held.

One battery of psychological experiments that test for the representational change faculty in young children is illustrated by the following: A child is shown a Smarties[6] box and is then asked, "What is inside the box?" The usual reply, "Smarties", indicates that the child has naturally come to believe that Smarties are in the box. Then the child is shown that the box actually contains a pencil. Once shown the pencil, the child is asked the question, "What did you think was in the box before you saw the pencil?" The answer "Smarties" is taken to indicate that the child has developed the cognitive capacity for representational change; the incorrect response, "a pencil," indicates the opposite. The children who answer incorrectly are thought to lack some aspect of the cognitive machinery necessary either to compute, represent, or report on their own previously held mental belief states when those states differ from currently held belief states. "They have no understanding of representational change. That is to say, they do not know that their beliefs have changed."[Astington and Gopnik, 1988] This particular experiment, and others like it, have been performed on children aged 3-5 years. The youngest of the experimental subjects tend to "fail" the test for representational change capacity, i.e., they answer "a pencil" to the second question. Subjects between 4 and

[5] "I used to think x, but now I *believe* y" seems more appropriate. Likewise the modality of knowledge should be replaced by that of belief in the false belief and ARD reports given below. The paraphrases for false beliefs and ARD reports also come from Astington and Gopnik. The paraphrase for pretense reports (also given below) is our own.

[6] "Smarties," the candy.

5 years of age do better, indicating a developmental acquisition of the cognitive capacity for representational change.[7]

False belief (not to be confused with "mistaken belief" as discussed earlier), as the term is used in the psychological literature (see [Wimmer and Perner, 1983]), the second related metacognitive task, differs from representational change in that the presumed erroneous belief is attributed to another person.[8] A paraphrase of a false belief report is "He thinks x, but I know y." The set of experiments that search for the false belief faculty and its development in children includes some that are very much like the representational change experiment just described. For instance, a child may be shown a box of Smarties and its contents, a pencil, and then asked what another child would think is inside the box if that other child were to be shown just the outside of the box. The child that responds, "a pencil", is taken to lack the false belief faculty as she is judged incapable of representing or computing another child's mental view of the world that contrasts with her own. On the other hand, the response, "Smarties", is taken to indicate a developed false belief faculty.

The third related metacognitive task, making the *appearance-reality distinction* (ARD) (see [Flavell *et al.*, 1986]) is evidenced by the report: "It looks like x, but really it's y." The canonical ARD illustration occurs when a white card is passed behind a red filter, making the card appear red to an observer.[9] Individuals able to make the ARD understand that the way something, be it an object or situation, looks or sounds (or more generally, appears) may differ from the way it is believed to be. Those that lack the ability do not accept or consider, as belief, the "correct" state of the world when appearances contraindicate that state.

Pretense, or pretend play, (see [Leslie, 1987]) is the last of the related metacognitive tasks to be considered here. Pretense is characterized by an intentional fabrication or distortion of facts, objects or events paraphrased by the report: "I pretend x, but I believe y." A child may

[7] Other plausible explanations need to be discounted. For example, that the children who fail this test have not yet fully grasped the past tense. They may understand the question "What *did* you think *was* in the box ..." as "What *do* you think *is* ...". (Bonnie Dorr, personal communication)

[8] There is also a temporal, or tense, aspect differentiating false belief from representational change that may be significant to the interpretation of the empirical results in psychology. Specifically, in the case of false belief, the incorrect view of the world and the correct view of the world, from the perspective of the agent of interest, are simultaneously taken to be accurate by different people. In representational change, the two views are believed to be correct only at different times.

[9] See [Maida, 1991] for an AI approach to a colored-card-under-various-lighting-conditions problem. Maida's work is akin to the psychological notion of *false belief* as discussed in the previous paragraph, from the perspective of an outside observer who maintains a mental model of some another reasoner's beliefs.

pretend that a toy tea set is real or that an empty tea cup is full; that a banana is a telephone or that an imaginary friend is his constant companion. But the child does not assent to the pretense outside of the stipulated pretend context.

1.5.2 Representation

An argument can be made for a connection among the four metacognitive tasks discussed above, namely performing each requires one to associate some (apparently) correct view of the world with some (apparently) incorrect view of the world. The surface expression of the reports associated with the tasks (i.e., "I used to think x, but now I know y," "He thinks x, but I know y," etc.) makes vivid this correct-view/incorrect-view pair and has been taken by some researchers as an indication that the four tasks share a common underlying cognitive representational scheme. Indeed, Leslie ([Leslie, 1987] and [Leslie, 1988]) has developed a representational theory, based on a correct world view/incorrect world view pair, that seemingly accounts for the four related tasks.

There is empirical evidence to suggest that this cognitive relationship exists between at least three of the four tasks. The evidence comes in the form of the above-mentioned near simultaneous acquisition of the tasks in children by the age of 3 to 5 years (see, for example, [Wimmer and Perner, 1983] and [Flavell *et al.*, 1986]). This evidence lends credence to the theory that the common representation scheme necessary for these tasks and its associated (cognitive) computational machinery are thought to be in place in children aged 3-5. The earlier acquisition of the fourth ability, pretend play, suggests further that the representation scheme itself may be in place in children by two years of age, but the mechanism needed to calculate (or report) the appropriate mental state attributions for ARD reasoning, etc., may not fully develop for another 1-3 years. That is, computation of pretend play mental states may be more elementary than those of the other metacognitive tasks. (See [Leslie, 1987] and [Leslie, 1988] for a detailed analysis of this view).

Another relationship among the four metacognitive abilities discussed here, one that is hinted at in the literature but to my knowledge has not been made explicit, concerns the form that incorrect world views may take. There seems no principled reason why any of the metacognitive tasks should differentially constrain the nature of their respective incorrect

world views. From any correct/incorrect ARD world view pair one ought to be able to form the basis of pretense, and vice versa. (Likewise for any other two of the tasks.) So for example, take any typical ARD scenario and give it a pretense twist: one can easily *pretend* that a white card is red without needing a red filter to pass in front of it. Similarly any standard pretense scenario, like pretending that an empty tea cup is full, can be imagined as an ARD story: deceptive appearances may fool one into believing that that a tea cup is full when, in reality, it is empty. It is worthwhile, then, to study reasoning that contrasts two views of the world. This observation is not in and of itself surprising, but it does have a practical consequence: a single taxonomy of incorrect world views can be used to characterize the nature of the mistaken beliefs that might be the foundation of any of the four tasks. One such taxonomy is provided in [Leslie, 1987] and [Leslie, 1988]. There three fundamental forms of pretense, each corresponding to a different cognitive state of affairs or incorrect view of the world, are identified. Each form, so Leslie argues, corresponds to one of three well known properties of attitude reports or sentences of mental state terms.[10]

The first type of pretense situation, called object substitution, is characterized by one's (cognitive) use of some object to stand in for a different object. For example, a child may pretend that a pillow is a dog, or a banana is a telephone.[11] The second form, attribution of pretend properties, occurs when one imputes pretend properties to an object or situation. For instance, a child may pretend that an empty tea cup is full. Creating imaginary objects, the third form of pretense, occurs when one invents and attributes characteristics to an imaginary object. For example, a child may pretend that she has an imaginary friend.

For Leslie's purposes, object substitution, attribution of pretend properties, and the creation of imaginary objects cover the range of incorrect world views. We will exploit the idea that Leslie's taxonomy of pretense can be generalized, and in chapter 3 use it to categorize what we are calling *mistaken beliefs*.

---

[10] We need not be concerned with the specifics of Leslie's discussion on this matter here. See [Leslie, 1987] for details.

[11] The banana/telephone example is taken from Leslie [Leslie, 1987] and [Leslie, 1988] as is the "full" empty teacup example.

Chapter 2

Belief: Background and Some Definitions

Belief is one of the so-called propositional attitudes or mental states (desire, fear, etc. are others) which one may direct toward some mental content. A particular type of mental state (i.e., believing, knowing, etc.) can be distinguished by the psychological function(s) it serves. Beliefs function as a basis for reasoning (inference) and, together with desire, form the underpinnings of agent planning, action, and intent.

It would be nice to discuss belief reasoning under the presumption that beliefs themselves are well understood but unfortunately that simply is not the case. Of particular concern is the uncertainty about the nature of belief representation – how beliefs are represented inside the "head" of a cognitive agent (see [Cummins, 1989] for a brief discussion of some possibilities).

One theory has it that beliefs are explicitly represented inside the agent's head in some mental language (or "Mentalese") [Fodor, 1979]. In AI it is not uncommon to make this representational assumption and to choose a first-order language (with quotation) to represent the content of agents' beliefs ([Haas, 1986], [Maida, 1991], and [Perlis, 1985]). Reasoning, or inference, under this scheme, is the mental manipulation of the syntactic objects that encode beliefs. This *belief representation assumption*[1] is made here.

The representational issue aside, we can now take a closer look at the nature of the belief systems we will consider, noting other assumptions along the way. Then, in chapter 3 we will examine some of the kinds of mistaken beliefs that these systems may come to believe they have held.

[1] This characterization has been adapted form Maida [Maida, 1992] though he calls it the *knowledge representation hypothesis*.

2.1 Interactive Belief Systems

Beliefs are embedded within belief systems which, in their simplest form, are taken here to comprise a *belief set* (often called a knowledge base) and a set of inference procedures which operate on the belief set. We shall consider a somewhat more complex system which we call an *interactive belief system* (IBS). (Alternately the terms "agent" and "reasoner" will be used to an IBS.) An IBS is a belief system which can interact with the external world through its *observation module* (to be described presently). Its belief set is updatable in "real-reasoning-time" in the following ways:

1. Beliefs may be acquired through "observation" – a term intended to include more than visual sensing – as the agent is *presented* with external data.

2. Beliefs may be acquired via inference performed upon the expressions representing (already held) beliefs.

3. Beliefs may withdrawn from the belief set.

Figure 2.1 shows a model of a rudimentary IBS.

2.1.1 Observation

The first means of belief update - observation - is taken to be this: data are sensed through the agent's sensory devices, sent to a recognition module, in which mental tokens representing worldly objects and concepts familiar to the agent are "selected" and subsequently used in the construction of perception-based beliefs.

An assumption here – which we can call the *background tokening assumption* – is that this selection process, indeed the entire recognition process, is not something over which the inference engine has direct control. Rather it is a background computational process. When an agent thinks he sees his car, an existing mental token of his car might be used directly in the construction of the appropriate beliefs. An alternative is to assume that some other new mental token is generated, which denotes the observed car, over which the inference engine can operate. Say this token is *percept_12* and say the agent's particular existing mental token for his own car is *mc* (for "my car"). Then, a reasoned view would cause the two to become

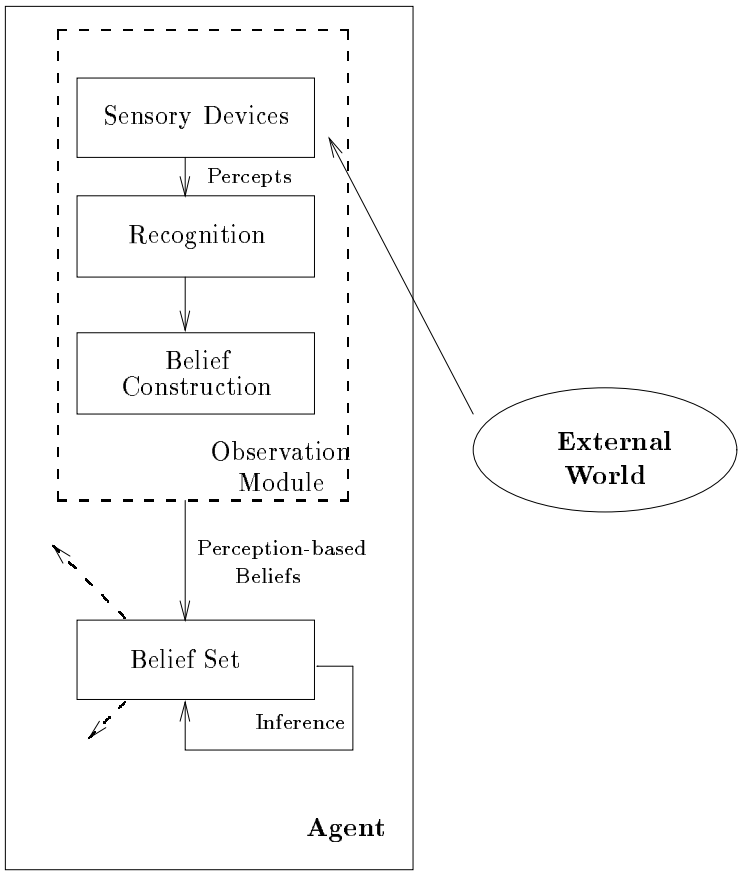

Figure 2.1: A rudimentary interactive belief system (IBS).

linked in the belief space through some sort of equality assertion, e.g.,

$$Eq(percept\_12, mc)$$

This latter approach is taken in [Maida, 1992]. A computational drawback is that it opens up the floodgates to a multitude of mental tokens that must be stored, evaluated, and reasoned about. When a person (or a robot) sees his car up ahead and walks toward it he is likely to look askance periodically to make certain that he doesn't bump into some other object along the way. As he moves, his viewing angle relative to the car changes, each potentially evoking a different percept token in Maida's sense. Are we to suppose that each time he glances back at his car he first simply sees a car-object, which he tokens as such, and then equates it to his car (and hence all other tokenings of his car)? Or does he adopt a view that there is always just one (relevant) car during the scenario?

The background tokening assumption does not imply that agents always correctly identify and token those objects which are presented to them, i.e., those objects to which their attention has been called; rather only that the identification, correct or incorrect, occurs at a "lower level" than inference. If and when an *object of presentation* is misidentified and the agent comes to see this, then a more appropriate token will, for the agent, come to be associated with that object of presentation. (The themes of *presentations* and *objects of presentation* will be addressed in more detail later.)

2.1.2 Inference

Beliefs acquired via the second means, mental manipulation of the contents of the belief set, rely on a set of inference procedures which, when applied, may alter the belief set.

A principle of Harman's (below) suggests what kind of inference procedures are embedded in belief systems.

> Recognized Implication Principle: One has a reason to believe P if one *recognizes* that P is implied by one's view. (italics added) [Harman, 1986] (p.18)

The key here is that commonsense reasoning is largely governed by inference procedures which rely on recognition, not only truth. That is, many facts may follow logically from our beliefs but the derivation of some are too complex or long for us to follow through. One way to see this is to identify Harmon's recognizability with a single-application version of modus ponens (MP). Here a reasoner only recognizes very short ("one-step") chains of implication. Suppose that a reasoner uses MP to immediately infer the proposition B from A and $A \rightarrow B$.[2] He does so because he is able to "recognize" that B is implied by the antecedent beliefs. However he may not in general use MP to immediately infer any β which logically follows from a set Γ of his beliefs. For instance, R may not follow directly, or "in one step", from P, Q, and $P \rightarrow (Q \rightarrow R)$. Instead an agent may first infer only $Q \rightarrow R$, because this one-step implication is recognizable. With this newly acquired belief available, he may subsequently use it as an antecedent for further inference. If Q happens to still be among the agent's belief set at that time, *then* he will recognize the implication and infer R. We will look more closely at this idea of step-wise reasoning in chapters 3–5.

[2] Intuitively, "immediately" can be taken to mean from one moment to the next. I will be more precise later.

2.1.3 Belief Retraction

Without belief retraction the fallible agent would forever remain quandried once she notices that her belief set is inconsistent or otherwise contains a mistaken belief. AI in general has seen the need for such non-monotonicity in reasoning systems.

Notice that once retracted, a belief is not necessarily gone forever. An agent may suppress a belief – retract it pending further evidence establishing its credibility or denial – and then reinstate it if and when that credible evidence becomes available.

Forgetting, denial, and amnesia are all unintended, or at least unreasoned, belief loss (retraction?) mechanisms. We will ignore these mechanisms and consider only reasoned (deliberate) belief withdrawal or retraction. Here belief retraction will typically be undertaken once a reasoned view suggests to an agent that the belief is somehow mistaken or no longer trustworthy. This might be due to an inconsistency, a faulty belief justification, or an observation-based misidentification error.

2.2 Beliefs and Time

We suggested that IBS's are situated in time. Let us be more precise about this and in turn more precise about what we consider current, past, and former beliefs.

We can use the ternary predicate symbol **Bel** to state that a proposition resides in an agent's belief set at a particular time. When we write the belief expression $Bel(a, \alpha, i)$ we intend a to denote a reasoning agent who holds the belief proposition denoted by (the quoted expression) α, at *attitude time* i. Expressions of the form $Bel(a, \alpha, i)$ are be called *belief reports*, as are English language renditions of such expressions. For example, $Bel(John, `tall(Mary)', t_1)$ says that at time t_1 John holds (held) the belief that Mary is (was) tall.

The idea behind attitude time is relatively straightforward: Time-situated reasoners have beliefs at various points in time, and it is often important to take account of these time points for a faithful representation of those beliefs.

The specific values that attitude time parameters can take may be determined by a clock external to the reasoning agent (e.g., days, seconds, hours) or they may be determined with respect to a reasoner's internal clock (e.g., the time it takes the reasoner to complete an

"inference step"). This latter approach is the one we take here, with inference steps denoted by non-negative integers.

Let a be a reasoning agent, α a wff, and t, t_1, and $t_2 \in \mathbf{N}$ denote reasoning steps (in time), then:

Definition 2.1 α is a *current belief* of a's at time t whenever $Bel(a, \alpha, t)$.

Definition 2.2 α is a *past belief* of a's at time t_1 if there exists a time $t_2 < t_1$ such that $Bel(a, \alpha, t_2)$.

Notice that a past belief at time t may also be a current belief at time t. Former beliefs differ from past beliefs in this aspect:

Definition 2.3 α is a *former belief* of a's at time t_1 if $\neg Bel(a, \alpha, t_1)$ and $\exists t_2 < t_1$ such that $Bel(a, \alpha, t_2)$.

In particular if it is *now* time t and $Bel(a, \alpha, t)$, then α is now current belief of a's. Similarly for a's past and former beliefs.

Another way to view this, and the one that will be used in chapters 3 - 5, is to look inside a given agent a's belief space as it evolves over time. Let

$$\mathbf{i}\text{: } \alpha,\ \beta,\ \ldots$$

denote that at time i the agent believes, perhaps among other things, α and β. Then α are β are current beliefs of our agent at time i, i.e., from the "outside looking in" we would assert both $Bel(a, \alpha, i)$ and $Bel(a, \alpha, i)$.

If a also believed γ at time $i - j$, for $i \geq j > 0$, i.e.,

$$\begin{array}{l} \mathbf{i\text{-}j}\text{: } \gamma,\ \ldots \\ \vdots \qquad\qquad \vdots \\ \mathbf{i}\text{: } \alpha,\ \beta,\ \ldots \end{array}$$

then, γ is a past belief of a's at time i, and if γ is not among those beliefs at time i (i.e., γ is different from α and γ is different from β, etc.) then γ is a former belief of a's at time i.

2.3 A Note on the Disposition of Beliefs

Harman considers three different dichotomies of belief that are worth mentioning in relation to *Bel* [Harman, 1986]. The three are: implicit/explicit, occurrent/dispositional, and conscious/unconscious.

An agent believes something explicitly if her "belief in that thing involves an explicit mental representation whose content is the content of that belief" [Harman, 1986] [pp. 13]. That is, an explicit belief is one whose content is "written" down in the reasoner's Mentalese. Implicit beliefs are not represented in this way, rather they follow from the agent's explicit beliefs in accordance with some principle(s) of reasoning.[3] For instance one may believe explicitly both that α and that $\alpha \rightarrow \beta$ yet believe only implicitly that β. *Bel*, as used above is intended to express an agent's explicit beliefs: $Bel(a, \alpha, t)$ if and only if α is an explicit belief of a's at time t.

A "conscious" belief, according to Harman, is one that an agent is aware of or can "easily become aware of simply by considering whether one has it" [Harman, 1986] (pp. 13-14). "Unconscious" beliefs are those that are not conscious. Both implicit and explicit beliefs, according to Harmon, can be either conscious or unconscious. *Bel* alone is not sufficient to distinguish beliefs along this dimension. His example of an explicit unconscious belief is this: "One might explicitly believe that one's mother does not love one, even though this belief may not be retrievable without extensive psychoanalysis" [Harman, 1986](p. 14).

Occurrent beliefs are those that are "either currently before one's consciousness or in some other way currently operative in guiding what one is thinking or doing" [Harman, 1986](p. 15). On the other hand, a belief that is not occurrent but is potentially occurrent is dispositional. Not all explicit beliefs are occurrent. Only those that are in the current "working set" are, the rest are dispositional. Consequently *Bel* alone is not sufficient to characterize a belief as occurrent or dispositional.[4]

In this work we focus on what Harman's explicit beliefs, particularly as they change in status from current to past, possibly to former, and possibly back to current again, etc.

---

[3] Levesque discusses this dichotomy formally [Levesque, 1984].

[4] The *memory model* described in [Elgot-Drapkin *et al.*, 1987] makes some headway into distinguishing occurrent from dispositional beliefs.

Chapter 3

Error: Background, Some Definitions, and Some Preliminary Analysis

> *Imagine what it would be like never to know that you had been mistaken, that you had held a false belief. This is not to say that all of your beliefs are true; you might be mistaken, but when you realize that your belief is false, you change the belief, and keep no record of your earlier belief.*[1] [Astington and Gopnik, 1988] (pp. 193)

Just what would it be like if, once a reasoner came to distrust a past belief, she could take certain appropriate actions to update her current set of beliefs, say by retracting the mistaken belief, but then had no recollection of how the mistaken belief related to her current belief set? We have already suggested some of the implications of such a limitation. Here we go into these issues in more detail so that we better understand what it takes for an agent to more fully correct her mistaken beliefs? We'll see that traditional TMS-style belief revision, even together with (mere) recall of former beliefs, is insufficient to account for all of the consequences of once having held the belief. Indeed, we usually know more about our former erroneous beliefs than this kind of historical information, and fortunately so. Specifically, we might come to know what was wrong with holding the belief – that is, what exactly about the belief was wrong – and why we came to acquire it in the first place. This information both aids in belief revision and is useful in explaining behavior based on the mistaken beliefs.

[1] Astington and Gopnik have not distinguished false from mistaken beliefs as I have. Their use of the term "false belief" in this quote can be interpreted either way, though I choose to read it as "mistaken belief".

3.1 Overview

A theme which will emerge concerns a reasoning agent's ability to exercise control of her own reasoning process, and in particular over her language. Traditional Tarskian semantical approaches to formal logic operate under the assumption that there is an a priori fixed domain of objects, each of which is granted the "property" of existence.[2] Objects are referred to in wffs of a given logical language by the use of a rigid designator [Kripke, 1980] (or set of designators) that is (are) to be used only and always to refer to that object. Thus a traditional formalism's ability to represent and construct wffs, is directly related to its associated Tarskian ontology.

In contrast, a language for belief reasoning (to be used in commonsense domains) must be flexible enough to reflect a reasoner's shifting ontological perspective as time passes [McCarthy and Lifschitz, 1987]. This is the case for natural language. For one, new terms denoting objects and concepts periodically enter and sometimes fade from languages in general and from a reasoner's own personal lexicon as well. Additionally the meaning of words change for us over the course of reasoning. And this is where error enters the picture.

One's ontology will be altered when she comes to learn the truth about Santa Claus, but the name "Santa Claus" need not be removed from her vocabulary simply because she comes to learn that he doesn't exist. Instead, the meaning of the term "Santa Claus" will change for her, reflecting that she now believes that he doesn't exist.

An agent may need to change her usage of an expression during the course of reasoning, and yet be able to recall the old usage and reason about both new and old in the current setting. An example that will receive attention in this chapter is the *Mistaken Car* problem from section 1.3. Our focus is on the incorrect use of a denoting term. ("Incorrectly" in a psychological sense, not in the strict referential sense.) When a reasoner sees what he thinks is his car, mc, in the parking lot space l he comes to believe $At(mc, l)$. If the car is not in fact his then, in a cognitively salient (but unwitting) sense, he has used the term mc to refer to two different cars; his and the car seen. He must come to see both usages in order to understand and correct his mistake.

Consider a more extreme example:

[2] Property is in quotes here in deference to the ontological argument – that existence cannot be predicated. See [Hirst, 1991] for a review of the difficulties surrounding existence in knowledge representation.

Agent 1: "Did you hear that John broke his leg?"

Agent 2: "No, really? That's a shame!"

Agent 1: "Yes, and his wife now has to do everything for him."

Agent 2: "Wife? John isn't married. Which John are you talking about?"

Agent 1: "I'm talking about John Jones."

Agent 2: "Oh, I don't know him. I thought you meant John Smith."

This apparently mundane conversation hides some very tricky features facing any formal representational and inferential mechanism, whether for use in natural language processing, planning, or problem-solving. For here occurs an implicit case of language control. As it dawns on the two speakers above that they are using the name "John" differently they need to reason about usage and adopt a strategy to sort out the confusion, e.g., by using last names, too.

The ability of a reasoning agent to exercise control of its own reasoning process, and in particular over its language, has been hinted at a number of times in the literature. Rieger seems to have been the first to enunciate this, in his notion of referenceability [Rieger, 1974], followed by others: [Perlis, 1985], [Perlis, 1988], [McCarthy and Lifschitz, 1987], etc. The underlying idea, as we conceive it here, is that the tie between linguistic entities (e.g., words) and their meanings (e.g., objects in the world) is a tie that the agent had better know about and be able to alter when occasion demands. This has a number of important commonsense uses, which have been listed elsewhere [Perlis, 1991].

A treatment is called for which allows rational behavior via a logic to be able to change usage, employ new words, and so on, much like our use of natural language. When a person newly discovers or learns about an object she might have no choice but to refer to it with an indexical or demonstrative expression like "that", "this", "he", "she", and so on. Alternatively her discovery might be occasioned by a new word (or word usage) or name that she can subsequently use to refer to the object.

The usual fixed language with a fixed semantics that is the stock-in-trade of AI and logic seems inappropriate to this task. In these traditional semantical approaches to logic, objects are a part of the logic's ontology only if the predetermined language includes a term to denote that object. This means that "newly discovered" objects, in the sense intended above, cannot

enter the ontology of an agent's reasoning using such a logic. In the the remainder of this chapter, I begin to develop the formal tools for reasoning about a particular type of perception-based error: compression. This development will carry over into the next two chapters. Here we propose "active logics" based on the step-logics of [Elgot-Drapkin and Perlis, 1990]. We apply these active logics to the specific issue of terminological change vis-à-vis mistaken beliefs.

3.2 Process

A reasoned change in belief can be viewed as a four stage process.[3] In the first stage the reasoner simply acquires (i.e., assents to) a collection of beliefs. She may continue to hold these beliefs for some time, acquiring new beliefs along the way, until some of these new beliefs lead her to notice a problem (perhaps a contradiction) with her current belief set. This is stage 2 of the process; the stage which initiates her reasoning about her mistake. In stage 3 she tries to tackle the particulars concerning her error, find which beliefs are troublesome, at least temporarily suspend the use of the troublesome beliefs as a basis for further inference, and try to discover what is wrong with the beliefs in question. Finally, in stage 4, she uses the information gathered in stage 3 to re-establish cognitive consistency, at least to the extent that she is unaware of any inconsistencies in her belief set. Here she is trying to revise her belief set to reflect her most recent, coherent view of the world. This may involve taking a more definitive stand on the previously suspended beliefs; some may be denied or rejected while others are reinstated. This may also involve some sort of modification to the beliefs in question. (We will see examples of this shortly.)

This process is sketched in figure 3.1. It is important to note that this process does not unfold in a temporal vacuum. Nor does the process occur as in a temporal logic where reasoning goes on in a timeless present about the past and future (see, for example, [McDermott, 1982]). Instead, there is a notion of an ever changing *now*, together with an ever evolving set of currently accepted beliefs, which provide the perspective from which to view erroneous (past) beliefs.

Moreover, the tidy encapsulation of the process as described here should not be taken to

[3] The *stages* here are not to be confused with the *steps* discussed informally in the previous chapter and more formally in the remainder of this work. It may take many reasoning steps to pass from stage to stage.

Stage 1:	Come to believe $\alpha_1, \ldots \alpha_n$ at some time t
Stage 2:	Notice a problem with time t's beliefs (inconsistency)
Stage 3:	Identify the problematic beliefs
	a. Find the α_i's to suspect and why
	b. Suspend belief in those α_i's found in stage 3(a)
	c. Initiate check to reject or reinstate each α_i
Stage 4:	Re-establish (apparent) consistency

Figure 3.1: Taking a view of one's mistaken past beliefs.

suggest that reasoning about mistaken beliefs is a process with a sharply delineated beginning and end. To the contrary, the process is continual and evolving. At times, even beliefs about erroneous beliefs may themselves eventually come to be viewed as erroneous.

3.3 Recognizing an Error

One first becomes aware of a mistaken belief during stage 2. But how? One (logic-based) approach to rationality suggests that we come to suspect an error upon noting competing or incoherent beliefs, in a current belief set. That an inconsistency might, momentarily, crop up during the course of commonsense reasoning should not be considered odd (though it should be considered a signal to re-assess one's beliefs). To the contrary, inconsistency seems almost a hallmark of ordinary reasoning. Unlike traditional omniscient formal logics, human reasoning is not deluged by the appearance of all possible wffs or beliefs once an inconsistency arises.[4] Instead of every assertion and its negation "swamping" one's set of beliefs, recognized inconsistency tends to remain confined to a few offending beliefs. Once one uncovers an inconsistency he might suspend the use of the offending beliefs (perhaps indefinitely) until he hits upon a way to settle the dispute. If he finds the fault then the belief set can be revised accordingly.

The process of coming to view a past belief as mistaken may be set in motion by a contradiction in one's currently held belief set. For example, at the moment that you try unsuccessfully

[4] Nor are the formal systems discussed in [Lin, 1987], [Priest and Routley, 1984], [da Costa, 1974], and [Elgot-Drapkin, 1988] swamped by all logical consequences of an inconsistency.

to unlock the door of a car that you have (mistakenly) taken to be your own, you may very briefly believe both that your key unlocks your car door (it does!) and does not unlock your car door (it does not unlock the door of the car you think is yours). This inconsistency may lead you to re-evaluate and try to sort out the confusion. You may think the key is damaged, or the lock is jammed, or you are using the wrong key. Regardless of the nature or outcome of these speculations, the momentary appearance and subsequent notice of the inconsistency sparked you into reasoning action.

One difficulty with using inconsistency to indicate a mistaken belief is that a full-blown consistency check of (the deductive closure of) a set of wffs (representing a reasoner's beliefs) is in general an undecidable affair. Assuming a computational model of cognition, such a test would place a stranglehold on real-time commonsense reasoning. Instead, we need a more limited test which heeds the observation that commonsense reasoners may be unaware of inconsistencies that logically follow from their beliefs, but if and when such awareness does set in, one will strive to do away with the problem by revising her belief set. To put it another way:

> Recognized Inconsistency Principle: One has a reason to avoid believing things one *recognizes* to be inconsistent. (italics added) [Harman, 1986](p.18)

3.3.1 The Step-logic Approach

Elgot-Drapkin and Perlis have developed formalisms, called step-logics [Elgot-Drapkin and Perlis, 1990] [Elgot-Drapkin, 1988], which offer a cognitively plausible solution to the problem of error detection via a decidable limited consistency check in the spirit of Harman's Recognized Inconsistency Principle.

Step-logic models reasoning as a one-step-at-a-time progression of inference and observation. Beliefs are represented at each step by a finite number of wffs. The limited test for consistency that Elgot-Drapkin and Perlis suggest amounts to scanning through the agent's current finite belief set for a wff and its negation, i.e., some α and $\neg\alpha$. This is the recognition part of Harman's principle; the reasoner is only "expected" to notice easy to recognize direct contradictions. Other inconsistencies may persist unchecked but in a relatively benign fashion because of the step-wise nature of the logic's inference mechanism. The idea here is this: one

might hold an inconsistent set of beliefs and not know it due to the sheer number or complexity of those beliefs. If it turns out that an easily recognizable direct contradiction results, then he'll take note and the appropriate corrective action. Unnoticed inconsistencies are not disastrous; the logics' step-wise inference rules control the inference process so that belief sets are not swamped with all of the logical consequences of an inconsistency, namely every wff.

Once the direct contradiction condition is met, other step-logic mechanisms (inference rules) can be invoked to address the avoidance aspect of Harman's principle. This includes the non-monotonic disinheritance or retraction of the contradictands (and other troublesome former beliefs).

Intuitively we can think of these logics as having time on their side. They afford the reasoning agent the time to try to sort out problems in her belief set and to modify the set in accordance with her most recently conceived view of the world.

In chapter 4 we will go into the technical details of step-logics. For now it is sufficient to illustrate the idea. In doing so we will make more precise the "one-step" chains of implication mentioned in section 2.1.2. The illustration involves our "step-at-a-time" version of MP. At step i the agent MP to infer β from α and $\alpha \rightarrow \beta$ if the antecedents are current beliefs at step i. The result of applying this rule is that β becomes a current belief at step $i+1$.

Now suppose that P, $P \rightarrow Q$, $Q \rightarrow R$, and $R \rightarrow \neg P$ (a deductively inconsistent set of beliefs) are all current at some time t. Then our agent will infer Q at step $t+1$, by "one-step" MP, but not R and not $\neg P$ (at step $t+1$). If all of the agent's beliefs persist from step t to $t+1$, then R will become a belief at step $t+2$ (from the step $t+1$ beliefs Q and $Q \rightarrow R$, using modus ponens). If this process continues then both P and $\neg P$, a direct contradiction, will appear at step $t+3$. At that point the direct contradiction can be noted and corrective action taken.

3.4 Error types

Inconsistency may indicate that a past belief is mistaken, but just what might the mistake be? Leslie's [Leslie, 1987] taxonomy (see section 1.5.2) offers a clue. A belief may mistakenly reflect an object misconception, the misattribution of a property, or the presumed existence of a non-existent object. For the remainder of this chapter we will mostly consider the first of

these and begin to develop the formal tools for reasoning about mistakes of this kind, which will carry over into the next two chapters.

3.4.1 Object Identification Errors: Compression and Dispersion

There are at least two ways in which the misidentification of an object may be reflected in one's beliefs. One is by "compressing" a denoting term and the other is by "dispersing" a set of denoting terms.[5] An instance of a compression-based identification error is characterized by a reasoner's use of a singly denoting term to refer to more than one object.[6] The cognitive flip side of compression is dispersion. A dispersion-based object identification error occurs when a reasoner mistakenly takes one object to be more than one object. This is reflected in an agent's belief set when she uses more than one singly denoting term, terms that she thinks to denote different objects, to refer to a single object. Let us now look at compression and dispersion in a bit more detail.

Compression

Our *Mistaken Car* story illustrates a compression-based error of misidentification. Let mc be the mental token that the reasoner in this story uses to denote his car. When he spots the car he believes to be his in the parking lot, say in spot l, then he will come to believe $At(mc, l)$.

In this expression mc, unbeknownst to the reasoner, is serving a dual role by referring, in a sense, to two different cars; his own and the one in spot l. This is a psychological claim. The reasoner has mentally confused mc with another car and this confusion is reflected by his use of mc in a belief intended in part to be about the other car. The situation is depicted in figure 3.2(a). Here the solid arrow pointing to the agent's car indicates a referential use of the term mc. The dashed line leading to the mistaken car indicates a demonstrative use of the same term mc; the term is being used to pick out *this* car – the one that the agent is looking at.

Rectifying the situation requires the agent to mentally distinguish the two confused cars, which in turn requires that he note and distinguish both his referential and demonstrative uses of mc in $At(mc, l)$. To do so requires the use of a mental token different from mc, perhaps newly

[5] My use of the terms "compression" and "dispersion" is borrowed from Maida [Maida, 1991].

[6] Or the use of members of one equivalence class of singly denoting terms to refer to one object (see [Maida, 1992] and [Maida, 1991]).

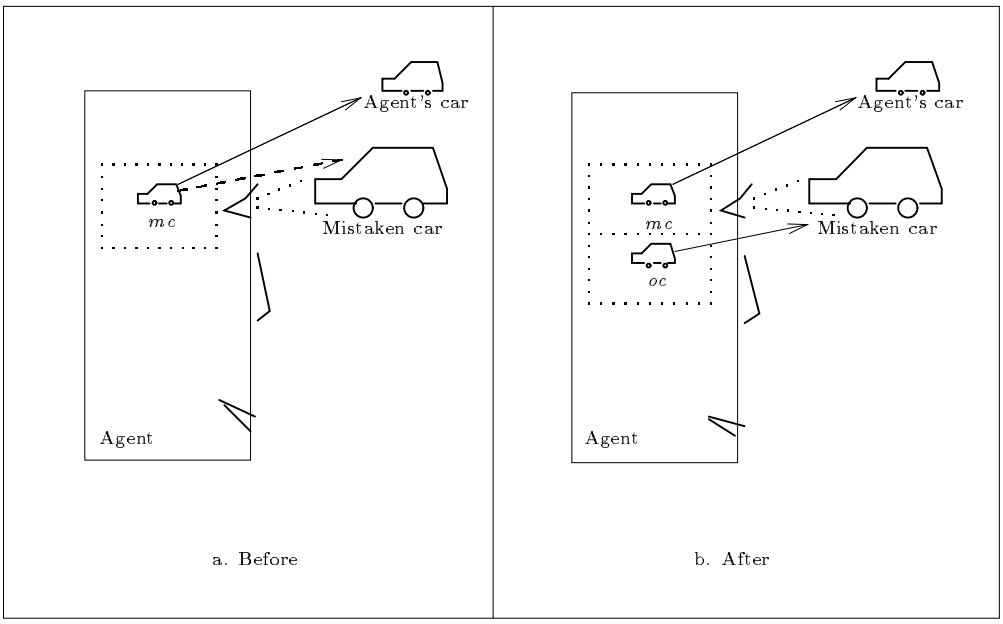

Figure 3.2: Mental representations of a compression-based identification error and its resolution.

created, which denotes (only) the car in the lot. This is depicted in figure 3.2(b) where *oc* (for "other car") is the new term used to pick out the mistaken car. (I will have more to say about the nature of the term *oc* shortly.) That *oc* and *mc* are mentally distinguished is indicated by the separate dotted boxes enclosing each. Recovering fully from the error involves revising his beliefs to reflect this newly discovered cognitive separation of the previously confused objects.

Dispersion

As mentioned a dispersion-based identification error is reflected in an agent's belief set by the use of two or more terms, thought by her to denote different objects, when they actually denote the same object. Consider the following illustration:[7]

> George's Car(s): You believe that George owns a blue Toyota. One day he drives to your house in a shiny red Toyota which you think is different from his blue car. Later you find out that he has repainted his car; the (red) car *is* his newly repainted (blue) car.

[7] Much has been written on what is perhaps the must famous philosophical example of dispersion: the identification of Hesperus (the evening star) and Phosphorus (the morning star) as distinct objects.

In this example your initial belief that the red car is different from the blue car is mistaken as it reflects a mental split of one object (the blue-turned-red car) into two. In making this error you have created anew a distinct mental token for an object, George's one and only car, that already had mental representation, and you have denied a correspondence between the two. This is depicted in figure 3.3(a) where each token is surrounded by its own box each with a referential arrow pointing to the same car. Correcting this error requires you to reshape the

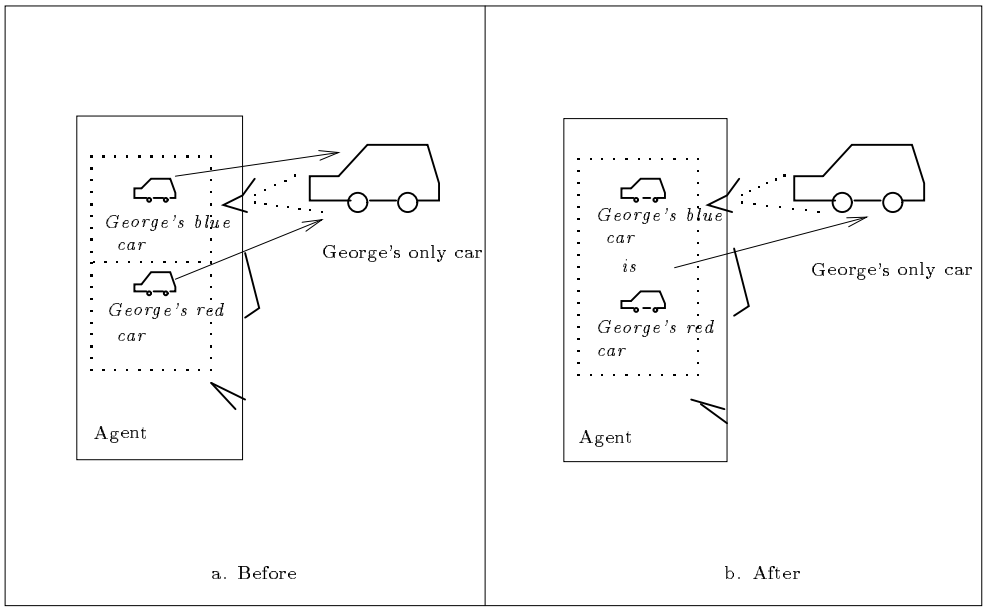

Figure 3.3: Mental representations of a dispersion-based identification error and its resolution.

relationship between the two tokens. There is no need to eliminate either token, in fact to do so may be counterproductive as you may want to refer to both (names) in explaining your error. Rather you must draw the tokens together by noting that they denote the same object. (See figure 3.3(b) – note the single box surrounding both tokens a the single arrow leading to George's car.)

The onset of dispersion is marked by the creation of a mental concept and associated name, say 'x', for an object, x, thought (incorrectly) to be distinct from another object, y, known by its own name, 'y'. In general, recovering from a dispersion-based identification error requires, in part, that a reasoner revise her beliefs to mirror a newly discovered cognitive merger of more that one token. This type of error is detected and corrected when the agent comes to

believe: (1) $x = y$ and (2) 'x' and 'y' are names that, as it turns out, were used, unknowingly, to refer to the same object.[8]

Denial versus distrust

The intent behind both the *Mistaken Car* and *George's Car* stories is that the reasoner eventually come to deny (i.e., assent to the negation of) his mistaken beliefs; $At(mc, l)$ is eventually denied in the former story and $george's\_blue\_car \neq george's\_red\_car$ is eventually denied in the latter. It is also possible, in fact it frequently occurs – often as a prelude to denying a belief – that one will come to suspect an error of compression or dispersion, say in light of a contradiction, without denying the belief. That is, she will distrust the belief (i.e., no longer accept it), and at the same time not assent to its negation.[9] Distrust without denial may occur fleetingly in the the *Mistaken Car* and *George's Car* stories – just prior to denial – but here is a more protracted example:

> The Twin(s): Imagine that you see a woman that you take to be your friend Kathy in a park throwing a softball with her left hand. Later you notice a woman, that you also take to be Kathy, batting right handed. Later still, you find out that Kathy has an identical twin, Patty. Believing that most (but not all) people throw and bat with the same hand, you come to wonder whether the softball tosser and the batter are the same person.

You have good reason to suspect an error of compression but without additional information you can not be certain. You have evidence weakly supporting each of two opposing views: (1) both "objects of presentation" looked the same and were in the park, so they may be the same person, but (2) most people throw and bat with the same hand, identical twins look alike but do not necessarily have the same dominant hand, so maybe it was two different people. The tension between the two views is sufficient to compel you to resolve the situation, but until you collect more evidence you remain uncommitted, distrusting but not denying your beliefs about Kathy's handedness.

[8] Maida addresses this problem in [Maida, 1992].

[9] Distrust will be made formal in chapter 5. So too will the notions of mistaken beliefs and perception-based misidentifications.

Maida has concentrated mostly on dispersion-based errors. We will attend to compression-based errors and their resolution for the remainder of this chapter (semi-informally) and in chapter 5 (formally).

3.5 Informal Stepped Reasoning: The *Mistaken Car*

Let us now take a closer look at the *Mistaken Car* story and see how it fleshes out in a somewhat informal step-like treatment which parallels the story as it unfolds in "real-reasoning-time". In what follows a series of reasoning steps that seem to be psychologically significant and relevant to that story are outlined. This treatment will be methodical, taking care to depict and discuss each step that we consider notable to our reasoner's evolving cognitive disposition. Each step i will have associated with it a number of beliefs which are intended to be the reasoner's relevant current beliefs at that time. Later, this progression will be formalized within a step-logic framework.

(In what follows, the term *mc* (for "my car") denotes the reasoner's car and *mk* (for "my (car) key") denotes the reasoner's key. Beliefs in the following figures are annotated to indicate how they (most recently) arose in the reasoning process. STORED means that the belief is among those held by the reasoner previous to his encounter with the car in the parking lot, REVISION means that a belief has been revised (resulting in the annotated one) after a mistake was uncovered, OBSERVATION indicates that the belief was introduced via the reasoner's observation module, and INFERENCE indicates that the belief was inferred via some, for now unspecified, inference rule. At each step after step 1, underlined wffs reflect beliefs newly acquired at that step.)

3.5.1 The Early Steps: Spotting the Car

To start, the reasoner may come to the parking lot with a slew of beliefs about his car: that it is blue, a Toyota, registered in Maryland, that his key fits it, and so on. Additionally, the reasoner may come to the scene with other commonsense knowledge, say that an object can not occupy more than one place at a time. Figure 3.4 illustrates this initial step of the reasoning process by depicting those previously held beliefs which we will concentrate on here.

Step 1:

$Registered(mc, maryland)$	[STORED]
$Color(mc, blue)$	[STORED]
$Make(mc, toyota)$	[STORED]
$Fits(mc, mk)$	[STORED]
$\forall xyz[At(x,y) \rightarrow (y = z \vee \neg At(x,z))]$	[STORED]

Figure 3.4: Initial beliefs in the *Mistaken Car* story.

Upon noticing and (mis)identifying the car in the lot the beliefs in figure 3.5 become current. In particular is the additional (ultimately viewed as mistaken) belief that mc is at location l, i.e., $At(mc, l)$.

Step 2:

$Registered(mc, maryland)$	[STORED]
$Color(mc, blue)$	[STORED]
$Make(mc, toyota)$	[STORED]
$Fits(mc, mk)$	[STORED]
$\forall xyz[At(x,y) \rightarrow (y = z \vee \neg At(x,z))]$	[STORED]
$\underline{At(mc, l)}$	[OBSERVATION]

Figure 3.5: Current beliefs relevant upon first noticing the car.

Notice two important features of the transition from step 1 to step 2. First, beliefs are held over, or *inherited* from step 1 to step 2. This is intended to model the persistence of a belief over time in the absence of a reason to suspend or retract that belief. Second, if we assume that no (psychologically salient) steps occur between steps 1 and 2 then, in particular, we are assuming that no mental name or token is generated which denotes the observed car other than mc over which the inference engine can operate. That is, the observed car is not pre-distinguished from mc up front and then inferentially equated or identified with mc once the reasoner (mis)identifies the car. This is the *background tokening assumption* of section 2.1.1.

To continue with the story, notice that an extended version of "one-step" MP, wherein variables are bound and substituted in the process of inference, e.g.:

$$\textbf{From } \forall x[(P(x) \rightarrow Q(x)] \textbf{ and } P(a), \textbf{ infer } Q(b)$$

will produce the belief $\forall z[l = z \vee \neg At(mc, z)]$ at step 3 from $At(mc, l)$ and $\forall xyz[At(x, y) \rightarrow (y = z \vee \neg At(x, z)))]$, both of which appear at step 2. We'll consider this to be a recognizable implication in Harman's sense. Step 2 is a snapshot of the reasoner's belief space just before he has made the inference and step 3; (figure 3.6) depicts his set of beliefs just after.

Step 3:

$Registered(mc, maryland)$	[STORED]
$Color(mc, blue)$	[STORED]
$Make(mc, toyota)$	[STORED]
$Fits(mc, mk)$	[STORED]
$\forall xyz[At(x,y) \rightarrow (y = z \vee \neg At(x, z))]$	[STORED]
$At(mc, l)$	[OBSERVATION]
$\underline{\forall z[l = z \vee \neg At(mc, z)]}$	[INFERENCE]

Figure 3.6: Inferring a new belief using extended MP.

To anticipate a bit, this newly inferred belief will turn out to be problematic because part of its justification, namely $At(mc, l)$, will turn out to be viewed as erroneous due to a a perceptually-based misidentification of the compression type. More will be said about this shortly.

3.5.2 The Middle Steps: The Key Doesn't Fit

When the agent tries to unlock the car door he fails. He tries to put his key in the lock, but it won't fit. He's using the correct key, or so he thinks. And it's his car, or so he thinks. But the key just won't fit and he comes to believe that his key does not unlock his car, however briefly. This belief, introduced at step 4 (figure 3.7), directly contradicts $Fits(mc, mk)$ which appears at step 3. Since there was no reason as of step 3 to retract, disinherit, or otherwise distrust $Fits(mc, mk)$, it will persist and it too appears at step 4. Step 4 then contains a

direct contradiction (a recognizable inconsistency in the Harman/Elgot-Drapkin/Perlis sense) reflecting the reasoner's confused state.

Step 4:

$Registered(mc, maryland)$		[STORED]
$Color(mc, blue)$		[STORED]
$Make(mc, toyota)$		[STORED]
$Fits(mc, mk)$	$\star$	[STORED]
$At(mc, l)$		[OBSERVATION]
$\forall xyz[At(x,y) \rightarrow (y = z \vee \neg At(x, z))]$		[STORED]
$\forall z[l = z \vee \neg At(mc, z)]$		[INFERENCE]
$\underline{\neg Fits(mc, mk)}$	$\star$	[OBSERVATION]

Figure 3.7: A contradiction occurs – the contradictands are starred ($\star$).

Again, to look ahead, it will turn out that $\neg Fits(mc, mk)$ is mistaken for much the same reason as $At(mc, l)$: an object identification error. But up to this point the reasoner is unaware of the specifics of his mistake. What he does become aware of immediately is that his beliefs are inconsistent, so *something* must be wrong, and something must be done about it. We will have the reasoner note the contradiction by using the binary predicate symbol **Contra**. $Contra(S, i)$ states that the components of its first argument, a set of beliefs, are direct contradictions which the reasoner simultaneously held at the step (or time) denoted by its second argument. For instance,

$$Contra(\{Fits(mc, mk), \neg Fits(mc, mk)\}, 4)$$

which appears at step 5 (see figure 3.8) states that the contradictands $Fits(mc, mk)$ and $\neg Fits(mc, mk)$ were both held at step 4.[10] (Note: At the point of noting a contradiction the reasoner has entered stage 2 as described in section 3.2.)

Since the contradiction has just occurred our reasoner has not yet had the time to reason through it, nor even time enough to speculate what might be wrong. Shortly he may stubbornly reject $Fits(mc, mk)$ and try again to jam the key into the lock, or he may re-examine the key to make sure it is the correct one, or he may take a closer look at the car to make sure it's

[10] For now the value of $Contra$ will simply be to mark that a contradiction occurred. Later, in chapters 4 and 5, it will serve a fundamental role for belief reinstatement.

his. But until then he has no reason to accept one of the contradictands and not the other. And yet both cannot be true so, pending further evidence, the agent should be unwilling to accept either. Another way to view this is that, for the moment, neither contradictand is to be trusted as a basis for further inference. (This is in line with Harman's principle of recognizable inconsistency from section 3.3.)

We use the binary predicate symbol **Distr** to express that a given belief, *Distr*'s first argument, is to be "distrusted" as of a particular step number, the predicate's second argument. (*Distr* will be made more precise in the next chapter.) Our reasoner will come to believe both $Distr(\neg Fits(mc, mk), 4)$ and $Distr(Fits(mc, mk), 4)$. Moreover, to ensure that use of the offending beliefs is suspended, they will be *disinherited* (in going from step 4 to step 5).

Figure 3.8 depicts these revisions. (*Contra*-ed and *Distr*-ed wffs are assumed to result from inference The details of a capable inference rule are discussed in chapter 4.)

Step 5:

$Registered(mc, maryland)$	[STORED]
$Color(mc, blue)$	[STORED]
$Make(mc, toyota)$	[STORED]
$\forall xyz[At(x,y) \rightarrow (y = z \vee \neg At(x,z))]$	[STORED]
$\forall z[l = z \vee \neg At(mc, z)]$	[INFERENCE]
$\underline{Contra(\{Fits(mc, mk), \neg Fits(mc, mk)\}, 4)}$	[INFERENCE]
$\underline{Distr(Fits(mc, mk), 4)}$	[INFERENCE]
$\underline{Distr(\neg Fits(mc, mk), 4)}$	[INFERENCE]

Figure 3.8: The contradiction is noted, distrusted and suspended.

The present example has been constructed so that $\neg Fits(mc, mk)$ will turn out to be judged erroneous (mistaken), and $Fits(mc, mk)$ will be reinstated. This because of the misidentification (observation-based identification error) that produced the former belief. But just how does this come to be known? We've already suggested that in some cases an agent might use a hypothesize-and-test process to try to ferret out the specific cause of his troubles from the set possible explanations that he can envision. A complete principled account of how one speculates and then confirms or denies her suspicions is beyond the scope of this thesis. It is likely that default reasoning is involved as is knowledge about the likelihood of errors (e.g., we

often haphazardly select the wrong key to try in a lock, popular cars are often misidentified since there are many similar looking ones, and so on) but the details will not be addressed here.

Instead, a simplifying assumption is to postulate a *tutor*, or friend, that can tell a reasoner about her errors. In this way we can avoid getting bogged down in issues of evidence, and the like and concentrate instead on the nature of erroneous beliefs and on belief revision. This is not as far fetched as it may seem at first. Imagine a friend saying to our reasoner, "Hey, this isn't your car!" Others note and help correct our errors frequently, and reliably. This is what we intend of the tutor. *Tutorials*, or statements which help to clarify or otherwise eliminate difficulties among the reasoner's belief set, are introduced like other observations at the appropriate step in the modeled reasoning process. (This use of tutorials is is in the spirit of McCarthy's advice-taker robot [McCarthy, 1958].)

We use the binary predicate symbol **Mstkn** to express that a given belief, $Mstkn$'s first argument, is considered to be mistaken in virtue of an (observation-based identification) error which was made at the step denoted by the predicate's second argument.

Notice that we are distinguishing the stance of distrusting a past belief from that of considering a past belief to be mistaken. The former stance is intended to reflect an agent's uncertainty vis-à-vis a past belief, especially in the face of contradiction, but does not imply any positive knowledge or evidence that the particular belief is in error. Thus, for instance, when a (direct) contradiction arises, and no other evidence is available to impugn the integrity of either contradictand, a rational agent may (temporarily) distrust both, and as a result suspend her (unqualified) acceptance and use of each. Viewing a past belief as mistaken is a stronger stance; it implies distrust, but it also implies that the agent has reason to suspect that the belief is in error, for instance that an object misidentification contributed to the the agent's originally holding the belief in question.

Notice also that neither of these stances, distrusting a belief nor viewing one as mistaken, implies the agent's denial of the belief in question, i.e., neither distrusting α nor considering it to be mistaken implies that the agent believes $\neg\alpha$. This is easy to see for the case of distrust which may consistently apply (simultaneously) to both $\neg\alpha$ and α, resulting in the suspension of both and the acceptance of neither. To see that considering some α to be mistaken does not

imply a belief in $\neg\alpha$ consider this: Suppose that I walk into an office looking for Sarah, whom I have spoken to over the phone but had never before seen, and identify the redheaded woman behind the desk as her. If I had no prior knowledge of Sarah's hair color then this incident is likely to lead me to believe that Sarah is a redhead, e.g., $Redhead(Sarah)$. If I later find out that I had indeed misidentified the redhead, i.e., she is not Sarah, then I should consider the belief $Redheaded(Sarah)$ to be mistaken – it was based on an object identification error – but the misidentification alone is insufficient for me to believe $\neg Redheaded(Sarah)$.

Our tutor will offer advice about mistaken beliefs. One simple kind of advice that a tutor can offer prescribes just which of the agent's observation-based beliefs are mistaken, e.g.,[11]

$$Mstkn(At(mc,l),5) \ \wedge \ Mstkn(\neg Fits(mc,mk),5) \tag{3.1}$$

(Shortly we will discuss more informative tutorials, but first we will use 3.1 to illustrate a weakness of such simple tutorials.

From tutorial 3.1, knowledge about the justifications (derivations) for his beliefs, and the appropriate inference mechanism the reasoner can revise his belief set in the style of traditional belief revision systems (e.g., TMS). Any mistaken belief will be distrusted (and retracted), i.e.,

$$Mstkn(At(mc,l),5)$$
$$Mstkn(\neg Fits(mc,mk),5)$$
$$Distr(At(mc,l),5)$$
$$Distr(\neg Fits(mc,mk),5)$$

Any belief previously justified, in part, by a mistaken belief will itself be considered mistaken, distrusted (and retracted), in particular:

$$Mstkn(\forall z[l = z \vee \neg At(mc,z)],5)$$
$$Distr(\forall z[l = z \vee \neg At(mc,z)],5)$$
$$Fits(mc,mk)$$

And some previously distrusted beliefs can be reinstated, i.e.,

[11] The tutorial depicted here is a single conjunctive wff which I take the liberty to immediately split into its two atomic components. The formal step-logic which I develop later will not include a rule of disjunction elimination, but one can easily be added.

$$Fits(mc, mk)$$

all at some future step (determined by the step-wise nature of inference mechanism, as it traces through belief justifications in the belief revision process).

Let us suppose further that the reasoner's object recognition model, or even the tutor, supplies the belief:

$$\exists x[x \neq mc \wedge At(x, l) \ \wedge \ \neg Fits(x, mk)]$$

i.e., there is a car different from mc at location l and mk does not fit mc. All of this done, the agent's beliefs are depicted in figure 3.9.

Step n:

$Registered(mc, maryland)$	[STORED]
$Color(mc, blue)$	[STORED]
$Make(mc, toyota)$	[STORED]
$\forall xyz[At(x, y) \rightarrow (y = z \vee \neg At(x, z))]$	[STORED]
$Contra(\{Fits(mc, mk), \neg Fits(mc, mk)\}, 4)$	[INFERENCE]
$Distr(Fits(mc, mk), 4)$	[INFERENCE]
$Mstkn(At(mc, l), 5)$	[TUTOR]
$Mstkn(\neg Fits(mc, mk), 5)$	[TUTOR]
$Mstkn(\forall z[l = z \vee \neg At(mc, z)], 5)$	[INFERENCE]
$Distr(At(mc, l), 5)$	[INFERENCE]
$Distr(\neg Fits(mc, mk), 4 : 5)$	[INFERENCE]
$Distr(\forall z[l = z \vee \neg At(mc, z)], 5)$	[INFERENCE]
$Fits(mc, mk)$	[REINSTATEMENT]
$\exists x[x \neq mc \wedge At(x, l) \ \wedge \ \neg Fits(x, mk)]$	[TUTOR]
$\neg At(mc, l)$	[INFERENCE]

Figure 3.9: The tutor "speaks" and beliefs are revised.

The reasoner's belief space is now consistent and it contains beliefs that reflect at least a partially accurate view of his world. There is even more information available to our reasoner at step n than might be available in some belief revision systems. In particular is the sort of historical information alluded to in Astington and Gopnik's quote reproduced at the start of this chapter, namely that the agent once held the mistaken beliefs. This information is extractable from the intended (still informal) semantics of $Contra$, $Distr$, and $Mstkn$. But still, the beliefs our agent holds at step n are insufficient for him to fully explain his situation.

Suppose that the car's owner comes along wanting to know why our agent was trying to get into her car. What response can the agent offer? Certainly not "I thought this car was mine", he does not have that information available. Moreover his current beliefs reflect that the the car in question is not his! So why was he standing there trying to put a key, a key that he now believes does not unlock the car, into the door of a car that he now thinks is not his own? The response "I was trying to steal your car" almost seems in order, but that is simply not true, he is not a thief – he believes that (though I haven't included this belief in the foregoing figures). Other possible responses seem inappropriate as well. The agent does not recall the car's owner asking him to try his key in her car door just for fun or curiousity's sake, so he will not respond with "You told me to try my key in your car door", and so on. But, given what he believes, almost any explanation is as plausible as the actual one; many different series of events could have led to this belief set. One reasonable partial explanation can be offered based on the (again, not shown) belief that the car looks like the agent's. But believing that a car looks like one's own and misidentifying that car as one's own are two very different things. There are many occasions when I have noticed a car that looks like mine and yet have not confused the two, in particular when I notice the look-alike while driving my own car.

Here is a explanation which is consistent with our agent's beliefs at step n and which is very much like the actual one, but is incorrect: he misidentified parking spot l instead of (or in addition to) misidentifying mc, and he misidentified mk – he tried the wrong key. Thus $At(mc, l)$ is mistaken not because the car was misidentified, but rather because the parking spot was. Likewise $\neg Fits(mc, mk)$ is mistaken because he mistook a different key to be mk.

Let us consider further the misidentification of l vs. the misidentification of mc. Both are depicted in in figure 3.10. If mc is the car shown in spot k in both drawings, then the reasoner will come to believe that his car is at location l both when he directs his attention to the car in space l, thinking it is mc (figure 3.10a) and also when he directs his attention to mc, thinking it is in spot l (figure 3.10b). Belief revision in the first case, the correct one according to our story, should be based on the fact that the thing he thought was mc, producing his beliefs $At(mc, l)$ and $\neg Fits(mc, mk)$, is not mc. Revision in the second case would be based upon the fact that the thing he took to be l is not l. (Likewise for the misidentification of the key.)

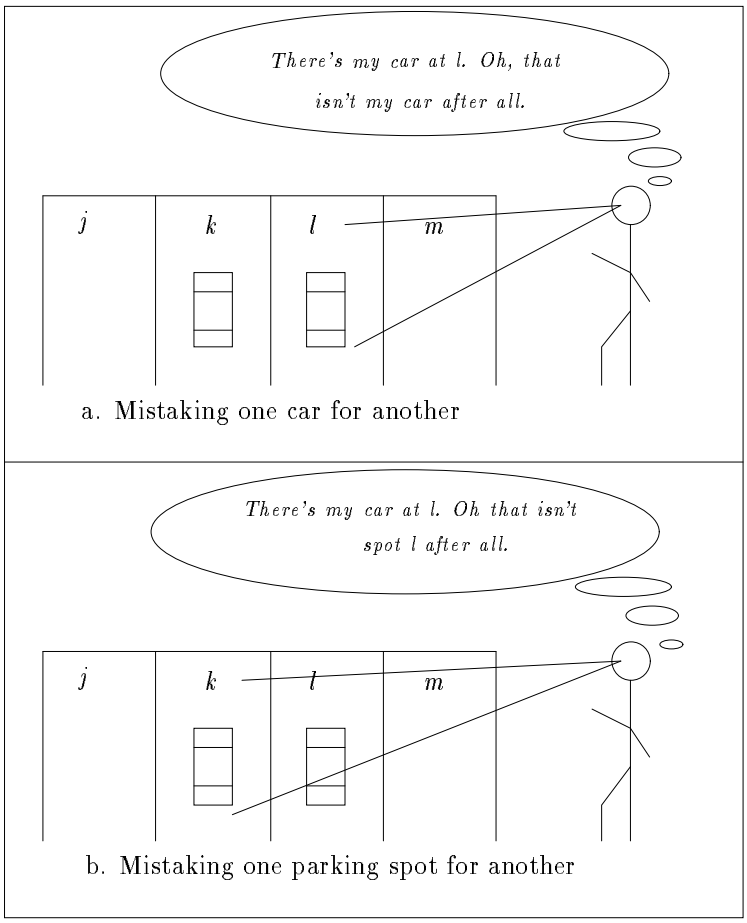

Figure 3.10: Two alternative models of underspecified tutorial.

The information necessary to distinguish these two cases is missing from tutorial (3.1). It concerns the changing meaning, for the reasoner, of the symbol *mc* as reasoning progressed from step 1 on. In the early steps, he thought he was using it only to refer to *mc*, but now it is time for him to learn that, in a sense, he also used it to refer to the other car. This is the issue we take up next.

3.5.3 The Later Steps: Presentations and *This* and *That*

A key informal idea here, one already alluded to several times, will be that of a *presentation*, which means roughly a situation or context in which attention has been called to a *presumed*

entity, but not necessarily an entity the reasoner has a very clear determination of at first.[12] This, we argue, is the case in virtually all situations initially, until we get our bearings. Before we actually make an identification we determine (perhaps unconsciously) that there is *something* for us to deal with. This is a small point as far as initial matters go, but becomes important if later we need to consider that something was wrong.

In the case of the *Mistaken Car* something, the reasoner tells himself later, made him think *this* (the car in location l) is *that* (his own). The something-or-other that brought about his mistake is what we call a presentation. Presentations will not play a formal role, but rather a motivational one in leading to our formal devices.

How can we formalize the notion of taking *this* for *that*? We begin by looking into the relationship between the two. Not a physical relationship, as in features that the two cars may share – though this may ultimately have a bearing on belief revision – but rather a cognitive relationship between the entities. This relationship, is suggested in the case of the mistaken car by the English statement, "I mistook *this* car to be *that* (my own)." The *this* here can be viewed as a demonstrative which (together with an appropriate demonstration) is used to pick out the mistaken car, the one in the lot. The *that* can be viewed as another demonstrative which is used to pick out mc. The statement, "I mistook this car to be my own", indicates a cognitive tie between two objects, automobiles in this case, that are in a sense linked in a (former) belief by the term mc.

Essentially what has happened is this: Upon first spotting the car our reasoner may be unaware of any recognition process, thinking simply that he sees mc. He is aware of an interest in one car only: his own. Then later, to comprehend or even suspect his mistake, he needs to become aware of an interest in two cars; his own (*that*) and the car he took to be his (*this*). In a sense, the term 'mc' in the original beliefs, $At(mc, l)$ and $\neg Fits(mc, mk)$ (as well as the inferred belief $\forall z[l = z \vee \neg At(mc, z)]$) refers to both of these cars. That is, the agent had his car in mind but connected his "mental image" of it to the wrong car, the one in location l. These beliefs reflect an unfortunate mental conflation, or compression, of these two cars that must be torn apart in the reasoning process. When two or more objects are cognitively compressed

[12] The vagueness in the notion of presentation does not, at this stage, hinder our formal treatment. However, we believe it will be necessary to clarify this notion. This is the focus of ongoing work. Among other things, it will involve a focus of attention, as hinted at by our informal "this" and "that" description below.

into one object, they are linguistically compressed in the representation and expression of one's beliefs. Even the suspicion of an error of compression requires, in part, the creation and introduction in the reasoner's mind of the appropriate mental tokens, one for the *this* and another for the *that*, which he will subsequently keep distinct.

We use the 4-ary predicate symbol **FITB** to state that an object of presentation is at first identified to be some (other) object, thereby producing a set of beliefs, i.e., $FITB(x, y, S, i)$ says that object of perception, x, which was presented at i, is at first identified to be y thereby producing the beliefs in S. Then we use expressions containing Russell's ι-operator to pick out the *this* that was (mis)identified as *that*. The idea behind ι-expressions is this: suppose $\alpha(x, y_1, \ldots, y_n)$ is an expression satisfied by exactly one object x, then the definite description "*the* $\alpha(u, y_1, \ldots, y_n)$" can be used to denote that object. For instance, since there is one dog sitting here with me I can use the phrase "the dog which is sitting here" to denote *this* dog. ι-expressions are formalized definite descriptions: we let $\iota x\alpha(x, y_1, \ldots, y_n)$ denote "the unique object such that $\alpha(x, y_1, \ldots, y_n)$", e.g., $\iota x[Dog(x) \wedge Sitting(x, here)]$ denotes the dog sitting here.

In the *Mistaken Car* scenario

$$\iota x FITB(x, mc, \{At(mc, l), \neg Fits(mc, mk)\}, 2) \tag{3.2}$$

can be used to denote "the unique object of presentation, presented at step 2, which was at first identified to be mc thereby producing the beliefs $At(mc, l)$ and $\neg Fits(mc, mk)$"; that is, *the car in the lot*. We call expressions like 3.2 *reality terms*; their function is to denote an object which was "presented" to an agent and (possibly) incorrectly identified by her in the past. Thus these are terms used to denote an entity, (possibly) replacing a previously held but incorrect description of that same entity. As a shorthand convention we use $tfitb(y, S, i)$, "the thing (object of presentation) which was at first identified to be ...", in place of $\iota(x)FITB(x, y, S, i)$. By incorporating reality terms we are able to express certain errors of object misidentification reflected in one's past beliefs. As an example,

$$tfitb(mc, \{At(mc, l), \neg Fits(mc, mk)\}, 2) \neq mc \tag{3.3}$$

is a tutorial asserting the appropriate error in the *Mistaken Car* story: "the unique object of presentation which was at first identified to be mc at step 2, thereby producing the beliefs

$At(mc, l)$ and $\neg Fits(mc, mk)$, is not mc. (We abbreviate assertions of the form $tfitb(t, S, i) \neq t$ by $MISID(t, S, i)$.)

Tutorial (3.1) should be viewed as a consequence of (3.3). The latter not only implies the mistakes stated in the former, it also expresses what that mistake was.

A virtue of tutorials like (3.3) is that they offer clues to belief revision beyond what appeared in figure 3.9. This is because the reality terms they contain name the misidentified object. Assuming that the mistaken beliefs are otherwise error-free, a plausible inference results when this newly created name is substituted for the misidentified term which appeared in the mistaken belief(s). (A formal substitution-based inference rule which accomplishes this is detailed in chapter 5.) Thus, asserting 3.3 sets in motion a belief revision process which is characterized, in part, by the following: the earlier beliefs $At(mc, l)$ and $\neg Fits(mc, mk)$ are disinherited or retracted, and both

$$At(tfitb(mc, \{At(mc, l), \neg Fits(mc, mk)\}, 2), l)$$

and

$$\neg Fits(mc, \{At(mc, l), \neg Fits(mc, mk)\}, 2)$$

are produced.

Figure 3.11 depicts our reasoner's relevant beliefs once tutorial (3.3) is asserted, and appropriate belief revision rules of inference are applied. In the figure we abbreviate the reality term

$$tfitb(mc, \{At(mc, l), \neg Fits(mc, mk)\}, 2)$$

by *oc* as a matter of convenience. A step number has not been specified here as it is assumed that several steps will pass from the moment the tutorial is asserted or introduced to get to this point. This should be viewed as the reasoner's current belief set once the last of those steps has occurred.

Figure 3.11 reflects the cognitive split we desire between the two cars in the story, *mc* and *oc*, as indicated by the following:

a. *mc* is registered in Maryland, it is blue, and is a Toyota

b. *oc* is at l, *mc* is not at l

$Registered(mc, maryland)$	[STORED]
$Color(mc, blue)$	[STORED]
$Make(mc, toyota)$	[STORED]
$\forall xyz[At(x,y) \rightarrow (y = z \vee \neg At(x,z))]$	[STORED]
$Contra(\{Fits(mc,mk), \neg Fits(mc,mk)\}, 4)$	[INFERENCE]
$Distr(Fits(mc,mk), 4)$	[INFERENCE]
$MISID(mc, \{At(mc,l), \neg Fits(mc,mk)\}, 2)$	[TUTOR]
$Mstkn(At(mc,l), 2)$	[INFERENCE]
$Mstkn(\neg Fits(mc,mk), 4)$	[INFERENCE]
$Mstkn(\forall z[l = z \vee \neg At(mc,z)], 5)$	[INFERENCE]
$Distr(\neg Fits(mc,mk), 4)$	[INFERENCE]
$Distr(At(mc,l), 5)$	[INFERENCE]
$Distr(\forall z[l = z \vee \neg At(mc,z)], 5)$	[INFERENCE]
$Fits(mc,mk)$	[REINSTATEMENT]
$\neg Fits(oc,mk)$	[INFERENCE]
$At(oc,l)$	[INFERENCE]
$\forall z[l = z \vee \neg At(oc,z)]$	[INFERENCE]
$\neg At(mc,l)$	[INFERENCE]

Figure 3.11: A tutor introduces a reality term and belief revision results.

c. *mk* fits *mc*, but does not fit *oc*

d. *oc* is not the same as *mc* (though it was once thought to be)

The process of reasoning glossed here, beginning with steps 1-5 and ending with the beliefs shown in figure 3.11 represents an informal solution to the *Mistaken Car*. A formal counterpart to this solution is detailed in chapter 5.

3.6 *One and Two Johns*

Let us look very briefly at two other stories related to the *Mistaken Car*, in that compression-based perceptual errors are made, but which bear more heavily on the issue of the flexibility of language during the course of reasoning. The problems are based on the dialogue presented at the start of this chapter:

Our *One John* example is very similar to that of the *Mistaken Car*, but will help us in moving toward the third example below. Here we imagine that we are talking to Sally about

a third person, whom we initially come to identify as our friend John, merely in virtue of matching John to Sally's description of the person, or the context of the conversation, etc., but not in virtue of hearing Sally use the name "John". Later we find out it is not John, but someone else.

There is no appropriate (perceived) entity before us which has been misidentified as in the case of the mistaken car; rather it is an abstract entity, a someone-or-other, still an object of presentation, the person that Sally had in mind. There is *this* someone that has been taken to be *that*, John. Our formalism treats abstract (objects of) presentation(s) of this sort much like the case of the *Mistaken Car*.

Now let us extend this to the *Two Johns* case: We are in a situation in which we are presented with a notion of a person, whom we (come to) think is our friend John. Then we are led to believe that he has a broken leg and his wife has to do everything for him. Later we suspect that there is a confusion, that not everything we are hearing makes sense. (John, our friend, is not married.) Is Sally wrong? Or have we got the wrong person in mind? Now here is the twist: Sally starts employing the name "John" to refer to this person.[13] Perhaps she is talking about a different John. To even consider this option we need to be able to "relax" our usage so that "John" is not firmly tied to just one referent. And later when Sally says that she is talking about John Jones, not our friend, John Smith, we need a way to refer to the two entities without *using* the term $John$. We may continue to *mention* the name, but judiciously, as it is ambiguous.

We can try to employ the same formal strategy that the agent used above up to a point. Namely, we may initially come to suspect that

$$tfitb(john, \{(BrokenLeg(john), Married(john)\}, 2)) \neq john$$

which has the English reading: "the unique object of presentation which was at first identified to be John at step 2, thereby producing the beliefs $BrokenLeg(john)$ and $Married(john)$, is not John." But then once we hear Sally use the the name "John" to refer to the person with the broken leg, whom we now believe is not our friend John, more must be done – the name "John" must be disambiguated.

[13] The sequence of events here is different than that reflected in the dialogue at the beginning of this chapter. Specifically, Sally uses the name "John" here only *after* we come to think that she is talking about our friend John. In the full paper we also discuss another version, in which Sally uses the name "John" at the outset.

This is where we must exhibit control over our language and language usage. First the ambiguity must be recognized. That is, we must come to see that *this* and *that* share the same name. Once that is done, new terms should be created, each to unambiguously denote one of the two Johns.

Proper naming and the use of names is made explicit with the the predicate symbol **Names**. We write $Names(x, y, i)$ to state that x names object y which first came to be known (by the reasoner) at time or step i; this could be weakened to time $\leq i$, or time $\geq i$, etc., if the exact time is not known. Including the third argument is somewhat non-standard, though not without a commonsense basis. We usually have at least a vague idea of when we come to know about someone. We can think of $Names(x, y, i)$ as collapsing $IsNamed(x, y) \wedge FirstLearnedAbout(I, y, i)$, where I is intended to be the first person pronoun.

To make ambiguity precise the binary predicate symbol **Amb** is used to state that a name does not refer uniquely beyond a certain step. Axiom **AM** expresses this:

$$\mathbf{AM} : (\forall x)(\exists yzij)[(Names(x, y, i) \ \wedge \ Names(x, z, j) \ \wedge y \neq z \ \wedge \ i \leq j) \rightarrow Amb(x, j)]$$

It says that if two different objects share a name, then the name is ambiguous for the reasoner once he became aware of both objects.

Once an ambiguity arises, our reasoner will need to disambiguate any belief using the ambiguous term. We use $RTA(x, y, i)$ to state that object x is referred to as y prior to step i. In particular if $Names(x, y, j)$ then $RTA(x, y, k)$ for $k > j$, $trta(y, i)$ is used an abbreviation for:

$$\iota x RTA(x, y, i)$$

"the unique thing referred to as y prior to step i", itself a non-ambiguous reality term.

Figure 3.12 gives a brief sketch of the evolution of reasoning we have in mind. (A formal solution is presented in chapter 5.) Figure 3.12 gives a brief sketch of the evolution of reasoning we have in mind. In the figure we use M, B, j, and 'j to abbreviate $Married$, $BrokenLeg$, $john$, and '$john$ respectively. Also $j1$ is used to abbreviate the expression $trta('j, 2)$, i.e.,

$$j1 = \iota x RTA(x, 'j, 2)$$

namely " the unique thing referred to as 'john' prior to step 2", and $j2$ is used to abbreviate

the expression $tfitb(trta(`j, 2), \{M(j), B(j)\}, 2)$, i.e.,

$$j2 = \iota x FITB(x, \iota y RTA(y, `j, 2), \{M(j), B(j)\}, 2)$$

namely "the unique thing which was first identified to be the the unique thing referred to as 'john' prior to step 2, which produced the beliefs $Married(john)$ and $BrokenLeg(john)$ at step 2." Thus $j1$ and $j2$ are newly created names, one for each John. We have omitted $Distr$-ed and $Mstkn$ expressions in the figure for the sake of conciseness. (Here ellipses (...) indicate that *all* beliefs from the previous step are inherited to the current step.)

Step 1: $\underline{\neg M(j), Names(`j, j, -\infty)}, AM$

Step 2: $\ldots, \underline{B(j), M(j)}$
(**Sally:** "...his leg is broken and his wife...")

Step 3: $AM, \underline{Names(`j, j, -\infty), Contra(\neg M(j), M(j))}$
(**Agent:** "Impossible! He isn't married.")

Step 4: $\ldots, \underline{MISID(j, \{M(j), B(j)\}, 2)}$
(**Sally:** "You misidentified who I'm talking about.")

Step 5: $AM, \underline{M(tfitb(j, \{M(j), B(j)\}, 2)),}$
$\underline{B(tfitb(j, \{M(j), B(j)\}, 2))}$
(**Agent:** "So that's what's wrong.")

Step 6: $\ldots \neg M(j)$
(<Reinstate Marital Belief>)

Step 7: $\ldots, \underline{Names(`j, tfitb(j, \{M(j), B(j)\}, 2), 2)}$
(**Sally:** "I'm talking about John.")

Step 8: $\ldots, \underline{Amb(`j, 2)}$
(**Agent:** "Oh, they have the same name!")

Step 9: $AM, \underline{\neg M(j1), M(j2), B(j2),}$
$\underline{Names(`j, j1), Names(`j, j2),}$
$\underline{j2 \neq j1,}$
(**Agent:** "Now I've got it.")

Figure 3.12: Sketch of stepped-reasoning in the *Two Johns* story.

Beliefs at step 1 are those held before the agent's conversation with Sally and those at step 9 reflect an unambiguous account of the two Johns, one now denoted by $j1$ and the other

by $j2$, once the problem is sorted out. In between are steps whose beliefs reflect information acquired via the conversation with Sally (steps 2 and 7) and via her advice (step 4); steps whose beliefs reflect that problems have been noted (a contradiction is noted in step 3 and the ambiguity is noted in step 8); and steps reflecting disinheritance (going from step 2 to 3, and from step 5 to 6).

The indicated steps have the following intuitive gloss: (1) the agent believes that John is not married, and is named "John". Then (2) comes to believe his leg is broken and he is married. This produces a contradiction, noted in (3), so neither marital belief is retained. Advice is then taken that John has been misidentified (4) which leads to the retraction (disinheritance) of the belief that John has a broken leg (6). The agent learns that the 'other person' is named "John" (7), notes the ambiguity (8), and takes corrective action (9) by creating and incorporating the unambiguous terms $j1$ and $j2$, one for each John.

Chapter 4

Step-logics: A Formalism for Time Situated Reasoning

There are several notable features of the stepped approach to reasoning illustrated in the previous chapter which will need to be preserved in a formal device applied to the specific issue of reasoning about past mistaken beliefs. Most conspicuous, of course, is that the reasoning be situated in a temporal context. As time progresses, a reasoner's set of currently accepted beliefs evolves. Beliefs become former beliefs by being situated in an ever changing "now", of which the reasoner is aware.

Secondly, inconsistency may arise and when it does its effect should not be disastrous (in the sense of *ex contradictione quodlibet*; from a contradiction all is entailed) rather it should be controllable and remedial, setting in motion a fairly broad belief revision process, which includes belief retraction.

Finally, the logic itself must be specially tailored to be flexible or "active" enough to allow, even encourage, language change and usage change when necessary.

As a theoretical tool the general step-logic framework developed in [Elgot-Drapkin and Perlis, 1990] and [Elgot-Drapkin, 1988] is well suited to these desiderata. But no heretofore developed step-logic offers a broad enough belief revision inference mechanism to suit our needs. The framework itself admits to non-monotonicity, so one foundational block for belief revision is in place, but others are missing in all previously defined step-logics.

In this chapter we first introduce the formal step-logic machinery and then discuss some drawbacks of previously defined step-logics. The primary achievement here is a solution to the

belief revision problem which will become part of the logics developed to solve the *Mistaken Car* and *Two Johns* problems.

Since contradiction and conflict play a key mediating role in the reasoning we are trying to formalize we pay special attention to these concepts here. Earlier step-logic work had a way to *ignore* contradictions. But more is needed. Not only must we adjudicate between contradictands, we must also prevent earlier mistaken beliefs (revealed by contradiction) from infecting *future* reasoning. Conflicting beliefs, mistaken beliefs, and their consequences must be controlled, so as not to infect other beliefs indefinitely into the future.

Recovering from contradiction was broached in Elgot-Drapkin's work, but only in an ad hoc way. There a conjecture was formulated, to the effect that, under (unspecified) circumstances, a step-logic should be able to regain consistency from an initially inconsistent set of beliefs. Here we begin to make inroads, in a limited way. We develop new step-logics which under suitable conditions are shown to recover from *direct contradictions* and their consequences (our *dc-recovery* theorem). This amounts to importing much of a truth-maintenance, or belief revision, system *into* the logic, which then – unlike a usual belief revision system – operates *during and as part of* the ordinary reasoning of the logic. This means that world knowledge can be brought to bear on the truth-maintenance (belief update) process, and other reasoning need not be halted while the belief updating is occurring. We advance two postulates concerning commonsense reasoning, the *short-chain* and *lazy-corroboration* hypotheses, which keep in check the computational bookkeeping required by our dc-recoverable step-logics.

These additions to step-logics, namely the mechanisms enabling terminological change and those for *dc*-recovery, have allowed us to solve commonsense problems centered around object-identification error as we'll see in chapter 5.

(Sections 4.1 and 4.2 are largely a synopsis of material found in [Elgot-Drapkin, 1988].)

4.1 Step-logics: An Inference Mechanism

Traditional descriptions of non-monotonic reasoning envision non-monotonicity as a relationship *between* theories: From one theory certain theorems follow that do not follow from an enlarged theory augmented with additional information (axioms). However, this relationship is expressed only in the meta-theory; the usual object-language logics pay attention to behav-

ior only *within* a fixed deductively closed theory. On the other hand, "theory change" is the central feature of the step-logic formalism. Instead of going (in the meta-theory) from one closed theory to another closed theory, we have a single evolving theory-in-progress that has facilities for reasoning about that evolution as it occurs.

In brief, a step-logic models belief reasoning by describing and producing inferences (beliefs) one-step-at-a-time, where the time of reasoning is integral to the logic. Complicated reasoning made of many successive inferences in sequence takes as many steps as that sequence contains. Error, change of mind, change of language, and change of language-usage all are time-tempered in that they are appropriately characterized only with regard to a historical account of beliefs, language, and its usage. Step-logics' one-step-at-a-time approach to inference offers a natural account of such histories.

A particular step-logic is a member of a class of step-logic formalisms; each particular step-logic is characterized by its own *inference* and *observation* functions. One distinguishing feature of step-logics is that only a finite number of beliefs (i.e., theorems) are held at any given discrete time, or *step*, of the reasoning process. Thus we can view each step as a discrete moment in a reasoning process. This attribute of finiteness is an important feature which, among other things, permits a computationally decidable treatment of *self-knowledge* – both positive and negative introspection into previously held beliefs, and of (limited) consistency-checking.

Let α, β, and γ (with or without subscripts) be wffs of a first-order language $\mathcal{L}$ and let $i, j, k \in \mathbf{N}$. The following illustrates what a step in the modeled reasoning process of a step-logic looks like:

$$\mathbf{i}\text{: } \alpha,\ \beta,\ \gamma,\ \ldots$$

The above display represents the belief set of the agent being modeled at step i, i.e., if it is now step (or time) i then α, β, and γ are currently believed.[1] The ellipsis is meant to indicate that there may be *finitely many* other beliefs held at this step.

[1] Note, as alluded to earlier, that the *steps* in step-logic reasoning are distinct from the aforementioned *stages* of the process of reasoning about former beliefs. Much time (i.e., many steps) may pass between a reasoner's coming to accept a belief (stage 1 of the process) and his coming to view that belief as mistaken (stage 2). Likewise, much time may intervene between stages 2 and 3 of the process.

4.1.1 Inference and i-theorems

Any wff α that appears at step i is called an *i-theorem* (roughly, a belief at step i). That α is an i-theorem is denoted by $\vdash_i \alpha$. A wff becomes an i-theorem in virtue of being proven (inferred) at step i. Proofs are based on a step-logic's inference function, which extends the historical sequence of beliefs one step at a time. An inference function can be viewed as a collection of inference rules which fire in parallel at each step in the reasoning process to produce the next step's theorems. For every $i \in \mathbf{N}$, the set of i-theorems are just those wffs which can be deduced from the previous step(s), each using only one application of an applicable rule of inference.

Inference rules, in their most general form, adhere to the structure suggested by rule schema **RS** below.

$$\textbf{RS:} \qquad \begin{array}{ll} \mathbf{i-j}: & \alpha_{i-j_1}, \ldots, \alpha_{i-j_m} \\ \vdots & \vdots \\ \mathbf{i}: & \alpha_{i_1}, \ldots, \alpha_{i_n} \\ \hline \mathbf{i+1}: & \beta_1, \ldots, \beta_p \end{array}$$

where $i, j \in \mathbf{N}$ and $(i-j) \geq 0$. The idea behind schema **RS** is this: at any step of the reasoning process the inference of β_1 through β_p as $(i+1)$-theorems is mandated when all of α_{i-j_1} through α_{i-j_m} are $(i-j)$-theorems, and all of α_{i-j+1_1} through α_{i-j+1_r} are $(i-j+1)$-theorems, ..., and all of α_{i_1} through α_{i_n} are i-theorems.

To illustrate the inference mechanism of step-logics we consider some simple, yet useful, inference rules in which only the previous step's theorems serve as a basis for a given step's deductions, that is where the j in **RS** is equal to 0.[2]

The first rule that we shall look at, called MP (for modus ponens), mandates the inference of β, at step $i+1$, given that both α and $\alpha \rightarrow \beta$ are i-theorems. (We discussed this rule informally in the previous chapter.)

[2] We will focus on rules where this condition holds for now, though later we shall relax it as we will require rules of the more general form.

$$\begin{array}{ll} \mathbf{i}: & \alpha,\ \alpha \rightarrow \beta \\ \hline \mathbf{i+1}: & \beta \end{array} \qquad \text{Modus Ponens (MP)}$$

Suppose that P, T, $P \rightarrow Q$, and $T \rightarrow (Q \rightarrow S)$ are among the set of, say, 2-theorems, then one single application of MP will produce Q as a 3-theorem and another, independent single application of the same rule produces $Q \rightarrow S$ as a 3-theorem. From these two 3-theorems a third application of MP produces S as a 4-theorem (see figure 4.1). Notice that a chain of

---

$\vdots$ $\qquad$ $\vdots$

Step 2: $\qquad P, T, P \rightarrow Q, T \rightarrow (Q \rightarrow S)$

Step 3: $\qquad Q, Q \rightarrow S$

Step 4: $\qquad S$

$\vdots$ $\qquad$ $\vdots$

Figure 4.1: Using MP.

(dependent) applications of MP is required to produce S as a theorem at step 4 in figure 4.1; step-logics do not operate by chaining these applications together in a single step to produce S as a 3-theorem. Notice too that MP says nothing about the $(i+1)$-theoremhood of α and $\alpha \rightarrow \beta$, the rule's requisite i-theorems. They may turn out to be $(i+1)$-theorems or they may not; MP itself is neutral on the issue. (As is **RS** in general.)

Another rule, called INH for *inheritance*, addresses the issue of theoremhood persisting from one step to the next. INH mandates the appearance (i.e., inference) of a wff α at step $i+1$, if α is an i-theorem.

$$\begin{array}{ll} \mathbf{i}: & \alpha \\ \hline \mathbf{i+1}: & \alpha \end{array} \qquad \text{Inheritance (INH)}$$

The motivation behind INH is this: once a reasoner comes to hold a belief, she may continue to hold that belief over time. Incorporating INH into the example given in figure 4.1 results in the sequence of theorems illustrated in figure 4.2. (For ease in reading, figure 4.2 reflects

---

⋮	⋮
Step 2:	$P, T, P \rightarrow Q, T \rightarrow (Q \rightarrow S)$
Step 3:	$P, T, P \rightarrow Q, T \rightarrow (Q \rightarrow S)$ $\underline{Q, Q \rightarrow S}$
Step 4:	$P, T, P \rightarrow Q, T \rightarrow (Q \rightarrow S)$ $Q, Q \rightarrow S, \underline{S}$
⋮	⋮

Figure 4.2: Incorporating the rule INH.

---

our continued use of the convention of underlining newly introduced wffs. At each step the underlined wffs are those which have been proven using a rule *other* than INH.) The use of INH accounts for the reappearance of step 2's theorems at step 3 and of step 3's theorems at step 4. MP has the same effect as in the previous example. Notice that there are two, seemingly indistinguishable, proofs of both Q and $Q \rightarrow S$ at step 4 in this example, one based on an application of INH and the other based on an application of MP.[3]

Unqualified inheritance via INH would be fine if our goal were a monotonic step-logic, wherein once proven, a theorem continues to be reproven at all subsequent steps. But commonsense belief reasoning is very much non-monotonic. Theoremhood (i.e., holding a belief) is sometimes contingent on the reasoner's not holding other beliefs. To put it another way, the acquisition of new beliefs may cause the reasoner to dispel or distrust or *disinherit* a former belief. Step-logics have been designed to permit the disinheritance of beliefs in such cases. This is accomplished by stipulating conditions which regulate the applicability of inference rules, and in particular, of INH. For instance, that a rational agent will not knowingly per-

[3] We will address the issue of distinct proofs or derivations shortly.

sist in holding contradictory beliefs may be modeled by stipulating that the rule INH not be applied to any i-theorem α when the the agent believes that α (directly) contradicts another of his beliefs (i.e., $\neg\alpha$).[4] This is not to say that contradictands, once disinherited, will never reappear at a later step. One or both may, but we can expect that this reappearance will be based on some new derivation(s). (More will be said about disinheritance shortly.)

4.1.2 Observations

Earlier we stated that a step-logic is defined not only in terms of its inference function (and language) but also in part by its observation function. An observation function, Obs, is a mapping from the set $\mathbf{N}$ of natural numbers onto the power set of wffs in $\mathcal{L}$. For each $i \in N$, the *finite* set $Obs(i)$ is comprised of the so-called i-observations, and is intended to represent the beliefs acquired by the agent while interacting with her environment at time i. These "observations" can be thought of as non-logical axioms or facts which the agent acquires over time. Observations are proven at a step in accordance with Rule OBS:

$$\begin{array}{ll} \mathbf{i}: & \\ \hline \mathbf{i+1}: \quad \alpha & \text{If } \alpha \in Obs(i+1) \qquad \text{(OBS)} \end{array}$$

Recall from figure 4.2 that several theorems appeared at step 2 for no formal reason (i.e., they were not proven). An observation function which is consistent with the example presented there, and one which explains the appearance of those wffs in figure 4.2 is:

$$Obs(j) = \begin{cases} P, T, P \rightarrow Q, T \rightarrow (Q \rightarrow S) & \text{if } j = 2 \\ \emptyset & \text{otherwise} \end{cases}$$

4.1.3 SL_7

The particular family of step-logics that will mainly be employed in this work is called SL_7. The distinguishing feature of the logics in this family is that they embody mechanisms for representing self-knowledge, time-situatedness, and belief retraction. (For full technical details of these mechanisms see [Elgot-Drapkin, 1988].)

[4] See figure 4.3 for a set of rules which incorporates this idea.

An example of an actual inference function (called Inf_B) for an SL_7 step-logic corresponds to the rules given in figure 4.3 (adapted from [Elgot-Drapkin, 1988]).

Rule 1	$\frac{\mathbf{i}:}{\mathbf{i+1}: \quad \alpha}$	IF $\alpha \in OBS(i+1)$
Rule 2	$\frac{\mathbf{i}: \quad \alpha}{\mathbf{i+1}: \quad \alpha}$	INHERITANCE[a]
Rule 3	$\frac{\mathbf{i}: \quad \alpha,\ \alpha \rightarrow \beta}{\mathbf{i+1}: \quad \beta}$	MODUS PONENS
Rule 4	$\frac{\mathbf{i}:}{\mathbf{i+1}: \quad Now(i+1)}$	AGENT LOOKS AT CLOCK
Rule 5	$\frac{\mathbf{i}: \quad \alpha_1(t), \ldots, \alpha_n(t), \ \forall x[\alpha_1(x) \wedge \ldots \wedge \alpha_n(x)] \rightarrow \beta(x)}{\mathbf{i+1}: \quad \beta(t)}$	EXTENDED MP
Rule 6	$\frac{\mathbf{i}: \quad \alpha,\ \neg\alpha}{\mathbf{i+1}: \quad Contra(\{\alpha, \neg\alpha\}, i)}$	CONTRADICTION NOTED
Rule 7	$\frac{\mathbf{i}:}{\mathbf{i+1}: \quad \neg K(\beta, i)}$	NEGATIVE INTROSPECTION[b]

[a] Where $\not\vdash_i Contra(\{\alpha, \beta\}, i-1)$. Also where α is not of the form $Now(j)$.

[b] Where β is a closed sub-formula at step i but is not an i-theorem.

Figure 4.3: Inf_B; rules of inference for a step-logic in the family SL_7.

Rules 1, 2, and 3 have already been discussed. They are OBS, INH, and MP respectively. Note the qualification (a) on rules 2 and 3 (INH and Negative Introspection) which stipulates under what conditions the rules do (not) apply.

Rules 4 – 7 are new. Rule 4 is intended to model an agent's awareness of her time-situatedness. It makes use of the distinguished predicate symbol **Now**; $Now(i)$ expresses the agent's belief that "it is now time i." By rule 4, at every step i, $\vdash_i Now(i)$. Since the current

time is meant to change at each step, and since the agent is meant to be aware of that fact, it is counterintuitive to allow $Now(i)$ to be inherited from one step to the next. The stipulation (a) on the rule of inheritance (rule 2) accommodates this by insisting that nothing of the form $Now(i)$ be inherited.

Rule 5 is an extension of modus ponens in which the antecedent of the rule contains a number of wffs, one of which is of the form $\forall x[(\alpha_1(x) \wedge \ldots \wedge \alpha_n(x)) \rightarrow \beta(x)]$ and the others are instantiated α_j's (i.e., $\alpha_1(t), \ldots, \alpha_n(t)$, where t is an instantiated term of $\mathcal{L}$). The result of applying this rule is the appearance of $\beta(t)$ at step $i+1$.

Rule 6 is the mechanism whereby direct contradictions are noted. The predicate symbol $Contra$ (discussed in the previous chapter) is used to mark two simultaneously occurring, directly clashing wffs i.e., α and $\neg\alpha$. Notice that the rule INH (rule 2) does not apply to wffs marked by $Contra$.[5]

Rule 7 is an example of a step-logic default rule. This rule offers one approach to negative introspection. It says that when some wff appears as a proper sub-formula of an i-theorem but is not itself an i-theorem then deduce $\neg K(\beta, i)$ as an $(i+1)$-theorem. K can be thought of as a knowledge predicate so that $\neg K(\beta, i)$ is intuitively read as " β was not known at step i." Since the set of theorems which appear at any given step is finite, the determination of the applicability of rules of this sort is computationally decidable. The condition that this rule should apply only to wffs which appear as sub-formulae at the previous step is simply one way to identify and constrain potential candidates over which the agent may negatively introspect. The thinking is this: Since the proper *super*-formulae are beliefs they are presumably "relevant" to the reasoner. So too then may be knowledge about any contained *sub*-formulae, including that they are not believed.

[5] Earlier treatments of $Contra$ (e.g., [Elgot-Drapkin, 1988]) considered it to be a ternary predicate in which the contradictands occupied the first two positions in its argument list and a step number occupied the third, i.e., $Contra_{elgot-d.}(\alpha, \beta, i)$. Contradictands were thus distinguished syntactically by their placement in the argument list, though the intended semantics of $Contra$ offered no principled reason to differentiate them in this way. The treatment here eliminates the syntactic quirk of handling $Contra_{elgot-d.}$'s first two arguments symmetrically.

4.2 An Example

In this section we will use the inference function depicted in figure 4.3 to briefly show an example of SL_7 in action. The purpose of this illustration is twofold. For one, we have only sketched the technical details of step-logics here, an example should help bring an intuitive grasp of the formalism within reach. Secondly, it will provide an opportunity to discuss several shortcomings of this, the heretofore most ambitious step-logic, that must be addressed before we proceed with the specifics of applying step-logics to reasoning about error and terminological change. (The example presented here is adapted from [Elgot-Drapkin, 1988].)

Let our logic be defined by Inf_B (figure 4.3) and Obs_B (below):

$$Obs_B(j) = \begin{cases} P \rightarrow B, P \rightarrow \neg F, \forall i[(B \wedge Now(i) \wedge \neg K(\neg F, i-1)) \rightarrow F] & \text{if } j = 1 \\ P & \text{if } j = k \\ \emptyset & \text{otherwise} \end{cases}$$

for a fixed $k > 1$. This observation function is an abbreviated version of the function that Elgot-Drapkin uses to solve a classic problem in the default reasoning literature regarding flying and non-flying birds. (More generally this problem illustrates the interaction between defaults and class-subclass hierarchies.) To see this interpret P as "Tweety is a penguin", B as "Tweety is a bird" and F as "Tweety flies". The observation function then tells us that our agent believes at step 1 that: (i) if Tweety is a penguin then Tweety is a bird; (ii) if he's a penguin then he doesn't fly; and (iii) the default that for all time steps i, if the current time is i *and* Tweety is believed to be a bird *and* the agent didn't know at time $(i-1)$ that Tweety doesn't fly *then* Tweety flies. Our agent observes no other facts until time k when he comes to believe that Tweety is indeed a penguin. Figure 4.4 is the sequence of steps which illustrate the inferences made by this step-logic. (In the figure the center dots $(\cdots)$ indicate a finite number of other beliefs which appear at a step but which have no bearing on the discussion that follows. For example both $\neg K(B,1)$ and $\neg K(P,1)$ are $(k-1)$-theorems which do not appear in the figure.)

Beliefs at steps 1 through $(k-1)$ are those held before the agent learns that Tweety is a penguin and those at step $(k+4)$ reflect an account of the story once a contradiction regarding Tweety's ability to fly has arisen and subsequently been resolved; Tweety is ultimately determined not to fly. In between are steps whose newly proven beliefs (those which are underlined)

Step 0: $\emptyset$

Step 1: $\underline{Now(1), P \rightarrow B, P \rightarrow \neg F, \forall i[(B \wedge Now(i) \wedge \neg K(\neg F, i-1)) \rightarrow F]}$

$\vdots$ $\qquad\qquad$ $\vdots$

Step k-1: $\cdots, P \rightarrow B, P \rightarrow \neg F, \forall i[(B \wedge Now(i) \wedge \neg K(\neg F, i-1)) \rightarrow F]$
$\underline{Now(k-1), \neg K(\neg F, 1), \cdots, \neg K(\neg F, k-2)}$

Step k: $\cdots, P \rightarrow B, P \rightarrow \neg F, \forall i[(B \wedge Now(i) \wedge \neg K(\neg F, i-1)) \rightarrow F]$
$\neg K(\neg F, 1), \cdots, \neg K(\neg F, k-2)$
$\underline{Now(k), \neg K(\neg F, k-1), P}$

Step k+1: $\cdots, P \rightarrow B, P \rightarrow \neg F, \forall i[(B \wedge Now(i) \wedge \neg K(\neg F, i-1)) \rightarrow F]$
$\neg K(\neg F, 1), \cdots, \neg K(\neg F, k-1), P$
$\underline{Now(k+1), \neg K(\neg F, k), B, \neg F}$

Step k+2: $\cdots, P \rightarrow B, P \rightarrow \neg F, \forall i[(B \wedge Now(i) \wedge \neg K(\neg F, i-1)) \rightarrow F]$
$\neg K(\neg F, 1), \cdots, \neg K(\neg F, k), P, B, \neg F$
$\underline{Now(k+2), F}$

Step k+3: $\cdots, P \rightarrow B, P \rightarrow \neg F, \forall i[(B \wedge Now(i) \wedge \neg K(\neg F, i-1)) \rightarrow F]$
$\neg K(\neg F, 1), \cdots, \neg K(\neg F, k), P, B, \neg F, F$
$\underline{Now(k+3), Contra(\{F, \neg F\}, k+2)}$

Step k+4: $\cdots, P \rightarrow B, P \rightarrow \neg F, \forall i[(B \wedge Now(i) \wedge \neg K(\neg F, i-1)) \rightarrow F]$
$\neg K(\neg F, 1), \cdots, \neg K(\neg F, k), P, B, Contra(\{F, \neg F\}, k+2)$
$\underline{Now(k+4), Contra(\{F, \neg F\}, k+3), \neg F}$

Figure 4.4: Step-logic $SL_7(Inf_B, Obs_B)$ in action.

reflect information acquired via: (i) observation (step k); (ii) (chained) inference using MP and extended modus ponens (steps $(k+1)$ through $(k+4)$); (iii) inference using the rule which notes contradictions (steps $(k+3)$ and $(k+4)$); and (iv) inference using the rule of negative introspection (steps $(k-1)$ through $(k+1)$). Also illustrated is disinheritance; in going from step $(k+3)$ to $(k+4)$ the beliefs F and $\neg F$ are both disinherited (though at the same time the latter belief, $\neg F$, is reproven, from P and $P \rightarrow \neg F$ using MP, and hence reappears at step $k+4$).

The indicated steps (beginning at step k) have the following intuitive gloss: (step k) the agent believes that Tweety is a penguin; (step $k+1$) since Tweety is thought to be a penguin the agent comes to believe that Tweety is a non-flyer, and also a bird, also the agent believes that he did not know that Tweety was a non-flyer in the previous step; (step $k+2$) a contradiction appears; (step $k+3$) the contradiction is noted, so neither flight belief is inherited to the next step; and (step $k+4$) that Tweety does not fly is reproven.

4.3 Addressing Some Shortcomings

The preceding introduction to step-logics should suffice to set the stage for a discussion of two shortcomings found in all heretofore developed step-logics which attempt to handle contradictions. Both of these failings must be addressed before we will be able to properly apply step-logics to model reasoned change in belief as prompted by mistaken beliefs. The problems we will consider here are (1) the inability of of previously developed logics to check, or halt, the lingering causes and consequences of contradictions, and (2) the inability of these same logics to reinstate observations as beliefs, under certain circumstances, after they have been disinherited.

4.3.1 The Lingering Consequences and Causes of Contradictions

Let's look at Inf_B from figure 4.3 (without loss of generality we will ignore the rule which introduces $Now(i)$ at each step i (rule 4)) and the observation function, Obs_1:

$$Obs_1(j) = \begin{cases} P, P \rightarrow Q & \text{if } j = k \\ \neg P & \text{if } j = k+n \\ \emptyset & \text{otherwise} \end{cases}$$

for fixed $k, n > 0$. Here P and $P \rightarrow Q$ will be (the only) k-theorems so by MP Q will be a $k+1$-theorem. Then (at step $k+n$) $\neg P$ is observed, causing a direct contradiction and the disinheritance of both P and $\neg P$. But Q persists, though its only derivation is questionable as it relies on P. (see figure 4.5). Thus, Q, a *consequence* of an untrustworthy theorem

---

⋮	⋮
Step k:	$\underline{P, P \rightarrow Q}$
Step k+1:	$P, P \rightarrow Q, \underline{Q}$
⋮	⋮
Step k+n:	$P, P \rightarrow Q, Q, \underline{\neg P}$
⋮	⋮
Step k+n+j:	$P \rightarrow Q, Q, Contra(\{P, \neg P\}, k+n)$
⋮	⋮

Figure 4.5: A belief (Q) based on a questionable former belief persists.

(belief) lingers beyond the step marking the disinheritance of its justification (P). Indeed, in this example Q will be inherited, and hence appear as a theorem, at *every* step $i > k+1$.

An even more pathological, though related, difficulty arises if we instead consider Obs_2:

$$Obs_1(j) = \begin{cases} Q, Q \rightarrow R, Q \rightarrow \neg R & \text{if } j = k \\ \emptyset & \text{otherwise} \end{cases}$$

In this example each of Q, $Q \rightarrow R$ and $Q \rightarrow \neg R$, which together with MP are the root *causes* of the contradiction R and $\neg R$, will persist indefinitely, and as a result so too will the contradiction (see figure 4.6.)[6] We can try to alleviate these problems by restricting the application of MP and INH. For instance, we might try this: at step i (1) INH should not apply to any α if $\vdash_i \neg\alpha$ (or, if α is of the form $\neg\gamma$, then INH does not apply if $\vdash_i \gamma$), and (2) MP should not apply to any α and $\alpha \rightarrow \beta$ if either of their direct contradictands (i.e., $\neg\alpha$ or $\neg\alpha \rightarrow \beta$) are i-theorems. The idea here is to (1) prohibit direct contradictands from being

[6] At the same time both R and $\neg R$ are also disinherited at each step beyond $k+2$ because of the stipulation (a) placed on INH in Inf_B.

$\vdots$ $\vdots$

Step k: $\underline{Q, Q \rightarrow R, Q \rightarrow \neg R}$

Step k+1: $Q, Q \rightarrow R, Q \rightarrow \neg R \underline{R, \neg R}$

Step k+2: $Q, Q \rightarrow R, Q \rightarrow \neg R, \underline{Contra(\{R, \neg R\}, k+1), R, \neg R}$

Step k+3: $Q, Q \rightarrow R, Q \rightarrow \neg R, Contra(\{R, \neg R\}, k+1), \underline{Contra(\{R, \neg R\}, k+2), R \neg R}$

$\vdots$ $\vdots$

Figure 4.6: The contradiction (R, $\neg R$) is reproven at each step.

inherited, and (2) restrict the use of MP to antecedent wffs whose contradiction is not also a belief.

Unfortunately these restrictions are insufficient to prevent the continual re-emergence of contradictions in certain cases. As long as the root cause of a contradiction persists, and no other action is taken, the contradiction will periodically re-arise (see figure 4.7, in which we again use Obs_2 and Inf_B augmented with our new stipulations).[7]

A solution to these problems must take into account the way inference is chained over the course of steps in step-logics. Any given i-theorem α may have been proven in any number of ways, where each distinct proof is based on (other) theorems appearing at previous steps. We can view α as the root of a proof tree whose nodes are the theorems used in deriving α and whose branches represent distinct proofs of α. The proposal offered here is to record the collection of wffs which appear on each branch of α's proof tree, along with α (at each step at which α appears), and use this information to (1) recover from the consequences of contradictions and (2) to prevent a contradiction from re-emerging.

4.3.2 *dc*-recovery: Some Preliminary Definitions

Let $SL(Inf, Obs)$ be an arbitrary step-logic with inference rules all of the form:[8]

[7] These new stipulations are nevertheless beneficial and will be used in the logic described shortly. (See Inf_{deriv}, figure 4.8.)

[8] The following definitions can be extended to apply to the more general rule schema **RS** discussed earlier.

$\vdots$ $\vdots$

Step k: $\underline{Q, Q \rightarrow R, Q \rightarrow \neg R}$

Step k+1: $Q, Q \rightarrow R, Q \rightarrow \neg R \underline{R, \neg R}$

Step k+2: $Q, Q \rightarrow R, Q \rightarrow \neg R, \underline{Contra(\{R, \neg R\}, k+1)}$

Step k+3: $Q, Q \rightarrow R, Q \rightarrow \neg R, Contra(\{R, \neg R\}, k+1)$

Step k+4: $Q, Q \rightarrow R, Q \rightarrow \neg R, Contra(\{R, \neg R\}, k+1) \underline{R, \neg R}$

Step k+5: $Q, Q \rightarrow R, Q \rightarrow \neg R, Contra(\{R, \neg R\}, k+1), \underline{Contra(\{R, \neg R\}, k+3)}$

$\vdots$ $\vdots$

Figure 4.7: The contradiction (R, $\neg R$) will alternately arise and then be disinherited.

$$\mathbf{k}: \ \underline{\quad \beta_1, \ldots, \beta_n \quad}$$
$$\mathbf{k+1}: \ \alpha$$

Definition 4.1 If $\vdash_{i+1} \alpha$ resulted from the application of an inference rule whose i antecedents are $\beta_1, \ldots, \beta_n$ then a *derivation set* (or simply, *derivation*) *of* α *at step* $i+1$ is a (possibly empty) set of theorems S containing exactly each of $\beta_1, \ldots, \beta_n$ and each wff in every derivation S_j (at step i) of β_j, for $1 \leq j \leq n$. (When a step number is understood we will simply say "derivation" instead of "derivation at step i". When we wish to call attention to the derivation S of α we write $\alpha[S]$.)

Note again that a theorem α may have more than one derivation at a step, each corresponding to a different branch in α's proof tree. For instance if MP is a rule of the logic and P, R, $P \rightarrow Q$, $R \rightarrow Q$ are all k-theorems then Q may have two different derivations at $k+1$; one including P and $P \rightarrow Q$ (and the theorems appearing in each of their respective derivations), and the other including R and $R \rightarrow Q$ (and the theorems appearing in each of their respective derivations).

Definition 4.2 Let $\vdash_i \alpha$, then α is *distrusted* at step $i+1$ iff:

(i) $\vdash_i \neg\alpha$ or if α is of the form $\neg\beta$ and $\vdash_i \beta$, or

(ii) $\exists\beta$ such that both $\vdash_i \beta[S_1]$ and $\vdash_i \neg\beta[S_2]$ and $\alpha \in S_1$ or $\alpha \in S_2$, or

(iii) each derivation of α at i contains at least one wff which itself is distrusted at step $i-1$.

We make our predicate symbol $Distr$ precise by using $Distr(\alpha, k)$ to assert that α is distrusted at step k.

Intuitively definition 4.2 says that an i-theorem is considered distrusted, or not trustworthy, at step $i+1$ if either (i) its negation is also an i-theorem, or (ii) it led to a direct contradiction, or (iii) each of its derivations contains a distrusted theorem.

Definition 4.3 $SL(Inf, Obs)$ *dc-recovers* (from the possible causes and consequences of all direct contradictions that appear at any step) if $\exists j$ such that $\forall k > j\ \neg\exists\alpha \vdash_{k+1} Distr(\alpha, k)$.

Definition 4.3 says this: a step-logic *dc*-recovers if there is a step j such that for any subsequent step k if α is a k-theorem then α will not be distrusted at step $k+1$.

Definition 4.4 $SL(Inf, Obs)$ is *eventually free of direct contradictions* if $\exists j$ such that $\forall k > j$ and $\forall\alpha$, either $\nvdash_k \alpha$ or $\nvdash_k \neg\alpha$.

Lemma 4.5 If $SL(Inf, Obs)$ dc-recovers then $SL(Inf, Obs)$ is eventually free of direct contradictions.

Proof: If $SL(Inf, Obs)$ dc-recovers then $\exists j$ such that $\forall k > j$ no k-theorem is $k+1$ distrusted by definition 4.3. Thus $\forall k < j, \alpha$ either $\nvdash_k \neg\alpha \nvdash_k \alpha$ by definition 4.2(i). Hence $SL(Inf, Obs)$ is eventually free of direct contradictions. ■

Theorem 4.6 $SL(Inf_B, Obs_1)$ does not dc-recover.

Proof: Illustrated in figure 4.5. ■

Theorem 4.7 $SL(Inf_B, Obs_2)$ does not dc-recover.

Proof: *Illustrated in figure 4.7.* ■

4.3.3 A New Step-logic

In this section we will develop a step-logic in the SL_7 family which does dc-recover given certain restrictions on its *Obs* function. (It will turn out that both Obs_1 and Obs_2 satisfy these constraints.)

We now introduce derivations formally into a new step-logic. Figure 4.8 depicts our new inference function, Inf_{deriv}.

Rule 1:	$\mathbf{i}: \quad$ $\mathbf{i+1}: \quad \alpha$		IF $\alpha \in OBS(i+1)$
Rule 2:	$\mathbf{i}: \quad \alpha[S]$ $\mathbf{i+1}: \quad \alpha[S]$		INHERITANCE[a]
Rule 3:	$\mathbf{i}: \quad \alpha[S_1], \alpha \rightarrow \beta[S_2]$ $\mathbf{i+1}: \quad \beta[\{\alpha, \alpha \rightarrow \beta(x)\} \cup S_1 \cup S_2]$		MP[b]
Rule 4:	$\mathbf{i}: \quad \alpha[S_1], \ \neg\alpha[S_2]$ $\mathbf{i+1}: \quad Distr(\alpha, i), Distr(\neg\alpha, i)$		CONTRADICTION DISTRUSTED
Rule 5:	$\mathbf{i}: \quad \alpha < S_1, \ldots, S_m >, Distr(\beta_1, i-1), \ldots, Distr(\beta_n, i-1)$ $\mathbf{i+1}: \quad Distr(\alpha, i)$		DISTRUST CONSEQUENCES[c]
Rule 6:	$\mathbf{i}: \quad \alpha[S_1], \neg\alpha[S_2], \beta[S_3]$ $\mathbf{i+1}: \quad Distr(\beta, i)$	$\beta \in S_1$ or S_2	DISTRUST ANTECEDENTS

[a] Where $\forall_i \ Distr(\alpha, i-1)$, $\forall_i \ \neg\alpha$, and for each $\beta \in S \ \forall_i \ Distr(\beta, i-1)$. Also, if α is of the form $\neg\gamma$ then this rule does not apply if $\vdash_i \gamma$.

[b] The stipulations placed on the antecedent $\alpha[S]$ of rule 3 (INH) in note (a) above apply to each of $\alpha[S_1]$, $\alpha \rightarrow \beta[S_2]$, and β here.

[c] Where each S_k contains at least one of $\beta_1, \ldots, \beta_n$ and α is not of the form $Distr(\gamma, j)$.

Figure 4.8: Inf_{deriv}

In the figure 4.8 the following abbreviations are used:

(1) α abbreviates $\alpha[\emptyset]$; i.e., we simply write α when α's derivation is the empty set.

(2) $\vdash_i \alpha < S_1, \ldots, S_n >$ if and only if $\vdash_i \alpha[S_1], \ldots, \alpha[S_n]$ and there is no S such that $\forall k$, $1 \leq k \leq n$, $S \neq S_k$ and $\vdash_i \alpha[S]$; that is $S_1, \ldots, S_n$ are *all* of α's derivations at step i.

(Since the limitations we will place on *Obs* (see the statement of the *dc*-recovery theorem, section 4.3.4) makes the derivation of any theorem of the form $Distr(\alpha, i)$ irrelevant, we annotate wffs of the form $Distr(\alpha, i)$ with $[\emptyset]$.)

Notice that derivations distinguish instances of theorems so that if $\vdash_i \alpha$ and α has multiple derivations at i, $S_1, \ldots, S_n$, then each of $\alpha[S_1]$,...,$\alpha[S_n]$ will appear as i-theorems.

The idea behind each of the rules of Inf_{deriv} is this:

Rule 1: (OBS) The derivation of an observation is empty indicating that no other beliefs have been used to derive it.

Rule 2: (INH) The derivation of an inherited belief is unaffected. Inheritance only applies to trustworthy beliefs: Namely, $\alpha[S]$ is inherited from step i to $i+1$ if it is not distrusted, its direct contradiction does not also appear at step i, and no $\beta \in S$ is distrusted. (See stipulation (a) in the figure.)

Rule 3: (MP) The derivation of a belief inferred via MP includes the wffs in the antecedent of MP (i.e., α and $\alpha \rightarrow \beta$) and all wffs contained in each antecedents' respective derivation. MP is applied only to trustworthy wffs as in rule 2 above. (See stipulation (b) in the figure.)

Rule 4: This rule marks a wff as distrusted at step $i+1$ when both it and its direct contradiction appear at step i. (Note: The predicate symbol *Contra* is not used here but it will return in the next chapter.)

Rules 5 and 6: These rules track down the consequences of *Distr*-ed beliefs (rule 5) and the antecedents of contradictory (distrusted) beliefs (rule 6). Rule 5 marks as *Distr*-ed at step $i+1$ any belief whose only derivations each contain a theorem distrusted at step $i-1$. (Notice that if *any* of an i-theorem's derivations contain an distrusted wff, those instances of the wff will not appear at step $i+1$ due to the stipulations placed on rules 2 and 3, regardless of the applicability of rule 5.) Rule 6 marks as *Distr*-ed any (antecedent) wff which appears in the derivation of a contradictory wff. That is, beliefs leading to a contradiction are themselves marked as distrusted.

A very simple example of Inf_{deriv} at work is based on the following observation function:

$$Obs_3(j) = \begin{cases} P, P \rightarrow Q, R, R \rightarrow Q & \text{if } j = 1 \\ \neg P & \text{if } j = 2 \\ \emptyset & \text{otherwise} \end{cases}$$

The resulting sequence of steps is shown in figure 4.9. Derivations are in **bold** type. Notice

---

Step 1: $\underline{P, P \rightarrow Q, R, R \rightarrow Q}$

Step 2: $P, P \rightarrow Q, R, R \rightarrow Q,$
$\underline{\neg P, Q[\{\mathbf{P}, \mathbf{P} \rightarrow \mathbf{Q}\}], Q[\{\mathbf{R}, \mathbf{R} \rightarrow \mathbf{Q}\}]}$

Step 3: $P \rightarrow Q, R, R \rightarrow Q,$
$Q[\{\mathbf{P}, \mathbf{P} \rightarrow \mathbf{Q}\}], Q[\{\mathbf{R}, \mathbf{R} \rightarrow \mathbf{Q}\}]$
$\underline{Distr(P, 2), Distr(\neg P, 2)}$

Step 4: $P \rightarrow Q, R, R \rightarrow Q,$
$Q[\{\mathbf{R}, \mathbf{R} \rightarrow \mathbf{Q}\}], Distr(P, 2), Distr(\neg P, 2)$

$\vdots$ $\vdots$

Figure 4.9: Inf_{deriv} at work.

---

the two instances of Q at step 2 each with a distinct derivation, one of which contains P which itself contradicts $\neg P$, also appearing at step 2. At step 3 the contradictands P and $\neg P$ are marked as distrusted and have not been inherited, though one derivation of Q at this step contains the distrusted contradictand P. This instance of Q, the one with P in its derivation, is disinherited at step 4 by stipulation (a) placed on INH which restricts inheritance to those instances of theorems containing no distrusted wffs in their derivations. By step 4 then, only one "clean" derivation of Q remains (and will continue to persist for all steps $i > 4$).

4.3.4 The *dc*-recovery Theorem

We now are prepared to prove that $SL(Inf_{deriv}, Obs)$ has *dc*-recovery when certain conditions apply to its observation function Obs. The following definitions help delineate those conditions.

Definition 4.8 An observation function Obs is *finite* if $\exists i$ such that $\forall j > i$, $Obs(j) = \emptyset$.

Definition 4.9 A wff α is *P-free* if α does not contain the predicate symbol P.

Definition 4.10 An observation function *Obs* is *P-free* if $\forall i \alpha$ if $\alpha \in Obs(i)$ then α is P-free.

Theorem 4.11 (***dc*-Recovery Theorem for** $SL(Inf_{deriv}, Obs)$) Let *Obs* be finite and *Distr*-free then $SL(Inf_{deriv}, Obs)$ *dc*-recovers.

The proof of theorem 4.11 uses the following lemmas and definitions.

Lemma 4.12 If $SL(Inf_{deriv}, Obs)$ *dc*-recovers then there exists a step RECVR such that $\forall k >$ RECVR, $\alpha \not\vdash_k Distr(\alpha, k-1)$.

Proof: By definition of *dc*-recovers. ■

Lemma 4.13 Let $SL(Inf_{deriv}, Obs)$ be as stated in the *dc*-recovery theorem, then $\forall k$ any k-theorem α containing the symbol *Distr* is of the form $Distr(\beta, j)$.

Proof: (By induction over k)

Base case: Let k=1. Only rule 1 can produce 1-theorems, all of which must be *Distr*-free by definition of *Obs*.

Inductive step: Assume the lemma holds for step n, then at step $n+1$ rule 1 can introduce only *Distr*-free theorems, again by definition of *Obs*; rule 2 reproduces an n-theorem which is of the appropriate form (by the inductive hypothesis); rule 3 does not apply by the inductive hypothesis; and rules 4 – 6 produce theorems of the appropriate form. ■

Corollary 4.14 Let $SL(Inf_{deriv}, Obs)$ be as stated in the *dc*-recovery theorem, then $\forall k, j, \alpha$ if $\vdash_k \alpha[S]$ then any $\beta \in S$ must be *Distr*-free.

Proof: Only rules 1, and 3 contribute new wffs to a derivation. This result holds by induction. Roughly, rule 1 must contribute a *Distr*-free wff since observations are *Distr*-free. Rule 3 only applies to *Distr*-free antecedents each with a *Distr*-free derivation (from the proof of lemma 4.13) thereby contributing only *Distr*-free wffs to the derivation of its consequent. ■

Definition 4.15 Let *Obs* be finite. Let OBSEND be the minimum step k such that $\forall j > k$ $Obs(j) = \emptyset$.

Lemma 4.16 Let $SL(Inf_{deriv}, Obs)$ be as stated in the *dc*-recovery theorem, then $\exists j \forall k > OBSEND$, *Distr*-free α, and S if $\vdash_k \alpha[S]$ then $|S| \leq j$.

Proof: Notice that for any step greater than OBSEND:

(1) Only rule 3 can increase the cardinality of a derivation S.

(2) Only rule 3 can contribute a *Distr*-free theorem which has never before appeared.

Notice also that if $\vdash_j \alpha[S]$ then for each $\beta \in S$ there is some step number, l, less than k such that $\vdash_l \gamma[S_j]$.

From step OBSEND on, only rule 3 can newly contribute as a theorem, and hence candidates for inclusion in a derivation, at most each sub-formula of every $OBSEND$-theorem (a finite number). Every derivation from step OBSEND on will contain at most this number of theorems plus the (finite) number of theorems which appeared before step OBSEND. ■

Definition 4.17 Let MAXLENGTH be the maximum cardinality of any derivation (of any theorem) in $SL(Inf_{deriv}, Obs)$ (as found in lemma 4.16).

Definition 4.18 Let MAXSTEP = OBSEND + MAXLENGTH + 1.

Lemma 4.19 Let $SL(Inf_{deriv}, Obs)$ be as stated in the *dc*-recovery theorem, then $\forall k > MAXSTEP$, *Distr*-free α, and S if $\vdash_k \alpha[S]$ then $\forall j$ such that $k > j >$MAXSTEP, $\vdash_j \alpha[S]$.

Proof: (Note that the only rule that can contribute a new wff to a derivation beyond step OBSEND is rule 3.) Suppose to the contrary. Then there is some step i, $k > i >$MAXSTEP, such that $\nvdash_{i-1} \alpha[S]$ and $\vdash_i \alpha[S]$. Thus there is some β such that $\vdash_{i-1} \beta[S_1]$, $\vdash_{i-1} \beta \rightarrow \alpha[S_2]$ (where $S = \{\beta, \beta \rightarrow \alpha\} \cup S_1 \cup S_2$) and neither β nor $\beta \rightarrow \alpha$ is distrusted at $i-2$. (Otherwise $\alpha[S]$ would not be an i-theorem by stipulations (a) and (b) on rules 2 and 3.) Either one or the other of β and $\beta \rightarrow \alpha$ is not a $i-2$ theorem, since (1) if both are and neither is $i-3$ distrusted then $\vdash_{i-1} \alpha$, which we assume not to be the case, and (2) if both are and either is $i-3$ distrusted then it would not be a $t-1$ theorem by stipulations (a) and (b) on rules 2 and 3. Without loss of generality, suppose it is $\beta[S_1]$ that is not a $i-2$-theorem. Then there is some $\gamma \in S_1$ (and hence $\gamma \in S$) which is not a $i-3$-theorem (for reasons analogous to those showing $\nvdash_{i-2} \beta[S_1]$). We can continue this backward progression, "removing" one

theorem from S at most MAXLENGTH times. But there are MAXLENGTH + 1 steps between OBSEND and MAXSTEP. ■

Proof: (of the *dc*-recovery theorem for $SL(Inf_{deriv}, Obs)$) By lemma 4.19 any *Distr*-free theorem which appears after MAXSTEP will also appear at MAXSTEP. Consider the set of all *Distr*-free $MAXSTEP$-theorems, $\alpha_1[S_1], \ldots, \alpha_n[S_n]$. Suppose rule 4 applies to any pair, $\alpha_i[S_i]$ and $\alpha_j[S_j]$ of these theorems (where $\alpha_i = \neg\alpha_j$). Then both of α_i and α_j will be disinherited at step MAXSTEP + 1 (by stipulations (a) and (b)), and never reappear by lemma 4.19. Thus no direct contradiction will appear after MAXSTEP, and neither rule 4 nor 6 will apply beyond this point.

If rule 5 applies beyond this point to some antecedent theorem α, then α will not appear at the next step (by stipulations (a) and (b)), nor at any subsequent step by lemma 4.19. There are a finite number of *Distr*-free wffs which can be forever disinherited in this way.

Thus rules 4, 5, and 6 will all cease to apply beyond some finite step, RECVR, and hence $\forall k >$ RECVR, $\alpha \not\vdash_k Distr(\alpha, k-1)$. ■

It is left for future research to characterize the set of theorems which survive the recovery process of $SL(Inf_{deriv}, Obs)$. Let this set be denoted by THM_{Obs}. We note that one characterization which does *not* apply to THM_{Obs} is this: *Let O be the set of all theorems introduced by Obs, and let M be a minimal subset of O whose complement $\overline{M}$ is consistent, then $\overline{M} \subseteq THM_{Obs}$*. To see that THM_{Obs} is not characterized in this way for every *Obs* let O be $\{P, \neg P\}$. Then $\overline{M} = \{P\}$ or $\{\neg P\}$. But notice that regardless of the step at which each of P and $\neg P$ is introduced via *Obs*, they will simultaneously appear at some step i. Thus they will both be disinherited at $i+1$, never to re-appear. Hence $\text{THM}_{Obs} = \emptyset$.

Computational Efficiency and Derivations

It might be argued that maintaining and searching through derivations at every step in the reasoning process is a computationally expensive task. This is true of reasoning that is comprised of long chains of inferences, as in mathematical reasoning; here derivations may be very long. It is also true of reasoning that relies on many simultaneous corroborations of the same hypothesis, as in some scientific reasoning; here there are many derivations of the same theo-

rem. But commonsense reasoning seems to be a different sort of process than both of these, one that is often (though not always) characterized by lots of world knowledge and rather (1) short chains of reasoning and (2) limited or lazy corroborations of beliefs.

One way to look at the "short-chain" hypothesis is that commonsense reasoners frequently touch base with reality, by getting external inputs, e.g., direct observation, testing, questioning, etc. Thus the reasoning gets regular validations or corrections, which can perhaps appropriately be treated a bit like new axioms. Of course, in commonsense reasoning axioms do not have a rigidly fundamental character as in mathematics, since we need to be able to account for error even in observations. Observations, then, may begin new chains of reasoning. Maintaining these short chains (derivations) has a computationally negligible effect.

The "lazy-corroboration" hypothesis asserts that we typically do not seek many independent corroborations or "proofs" (derivations) of our beliefs. This is not to say that we deliberately avoid corroborations, nor that we always feel content with just one or two. There are times when it becomes extremely important to secure as much evidence as possible before accepting a belief; say a plan to escape in a life and death situation. But, in general, we tend to readily accept beliefs and seek corroborations only as needed; we take a "lazy" approach to belief corroboration. As I look out the window and see what I think is my truck in the parking lot I simply believe that it *is* my truck. I don't have to go outside and try the key in the door, or check the vehicle's identification number, or peek through the windshield to see the empty coffee mug I left in there this morning to help verify that it is, indeed, my truck.

4.3.5 Reinstating Observations

We mentioned two shortcomings of previous versions of step-logics which need to be addressed here. The first, the lingering theoremhood of the causes and consequences of contradictions, is handled fairly generally by Inf_{deriv}. The second problem can be viewed as the dual of the first. Instead of being concerned with theorems persisting beyond their justifications, here we consider certain disinherited theorems which seemingly should be, but are not, reinstatable.

Consider the following dialogue:

Agent 1: "Where is your truck?"

Agent 2: "It's in the parking lot behind the building."

Agent 1: "No it's not. Take a look for yourself."

Agent 2: "Oh no! I parked it there this morning. Where is it?"

Let us use my_truck to denote Agent 2's truck and lot to denote the parking lot. We suppose that at some point Agent 2 will simultaneously, though briefly, hold the contradictory beliefs $Loc(my\_truck, lot)$ and $\neg Loc(my\_truck, lot)$. If we were to model her reasoning using either Inf_B or Inf_{deriv} both of these beliefs would subsequently be disinherited in virtue of the contradiction.

The intent of the above dialogue is that at its end Agent 2 will believe $\neg Loc(my\_truck, lot)$, though not because she re-observed her car missing from the lot. We would like a step-logic to allow a belief to be reinstated once the (partial) cause of its disinheritance, in this case its direct contradiction, is found to be mistaken, but no heretofore developed step-logic offers this reinstatement capability.

One proposal with some intuitive appeal is to develop a logic which operates on a *principle of recency*: that more recent observations be trusted, and hence not disinherited, when they conflict with earlier observations or otherwise inferred beliefs. The appropriate rules of inference could be altered to accommodate this, but as a general solution this won't do. The observation function is meant to model the agent's interaction with her environment, and appearances (i.e., beliefs constructed in accord with one's interaction with the environment) can be deceiving. Trust in recent observations may be a good initial response, or valuable in certain circumstances, but a truly robust logic should also be accountable to deceptive appearances, no matter how recent the observation which led to deception.

Consider the following more comprehensive approach: Suppose that an agent comes to note a contradiction between two beliefs. Let her initially distrust, and hence disinherit, both but allow her the further opportunity to note later that one or the other was mistaken for some known reason. If she does acquire this additional knowledge (about only one of the contradictory beliefs) then she can reinstate his faith in the other.

To formalize this approach we reintroduce the binary predicate symbol $Mstkn$ (originally discussed in chapter 3), which like $Distr$, takes as its first argument a (quoted) wff and as its second argument a step number. Intuitively, $Mstkn(\alpha, j)$ states that the wff α is mistaken in

virtue of something (e.g., an identification-based error) which occurred prior to or at step j.

We will limit the scope of negative introspection (rule 7 from inf_B) by insisting that a *Distr* wff be its only trigger. Strictly speaking this modification is not necessary, but it makes clear the idea being pursued here: that often it is useful for an agent to have available the fact that she does not know that certain distrusted beliefs are mistaken.

$$\mathbf{i}: \quad \frac{Distr(\alpha, j)}{\mathbf{i+1}: \quad \neg K(Mstkn(\alpha, j)i)} \qquad \text{where } \nvdash_i Mstkn(\alpha, j) \qquad \text{NEGATIVE INTROSPECTION}$$

Then we add a rule to handle the actual reinstatement of certain disinherited beliefs, specifically, the one of a pair of contradictands that is not know to be mistaken.

$$\mathbf{i}: \quad \frac{\begin{array}{l} Contra(\{\alpha, \beta\}, k), \\ Mstkn(\beta, j), \\ \neg K(Mstkn(\alpha, k), i-1) \end{array}}{\mathbf{i+1}: \quad \alpha} \qquad \text{REINSTATEMENT}$$

We now have the tools needed to solve the *Missing truck* problem. We use rules 1, 2, and 4 from Inf_{deriv}[9] (the others have no effect on this example) and the two rules above. The observation function, Obs_{truck}, is:

$$Obs_{truck}(j) = \begin{cases} Loc(my\_truck, lot) & \text{if } j = t_1 \\ \neg Loc(my\_truck, lot) & \text{if } j = t_2 \\ Mstkn(Loc(my\_truck, lot), t_2) & \text{if } j = t_3 \\ \emptyset & \text{otherwise} \end{cases}$$

where $0 < t_1 \leq t_2 \leq t_3$.[10] The sequence of steps shown in figure 4.10. (To condense the figure an i-theorem, of the form $\neg K((\alpha, l), j : k)$, where $j < k$, is used to abbreviate that each of $\neg K((\alpha, l), j)$, $\neg K((\alpha, l), j+1)$, ..., $\neg K((\alpha, l), k)$, are i-theorems.) Notice here that $\neg loc(my\_truck, lot)$ is disinherited at step $t+2$ because it contradicts $loc(my\_truck, lot)$, but since the latter turns out to be $Mstkn$ the former is reinstated at step $t_3 + 1$.

[9] Actually we use rule 4 from Inf_{deriv} along with the additional consequent $Contra(\{\alpha, \neg\alpha\}, i)$ as in rule 6 of Inf_B.

[10] Only the relative order of the introduction of these beliefs is important here, not the actual step numbers.

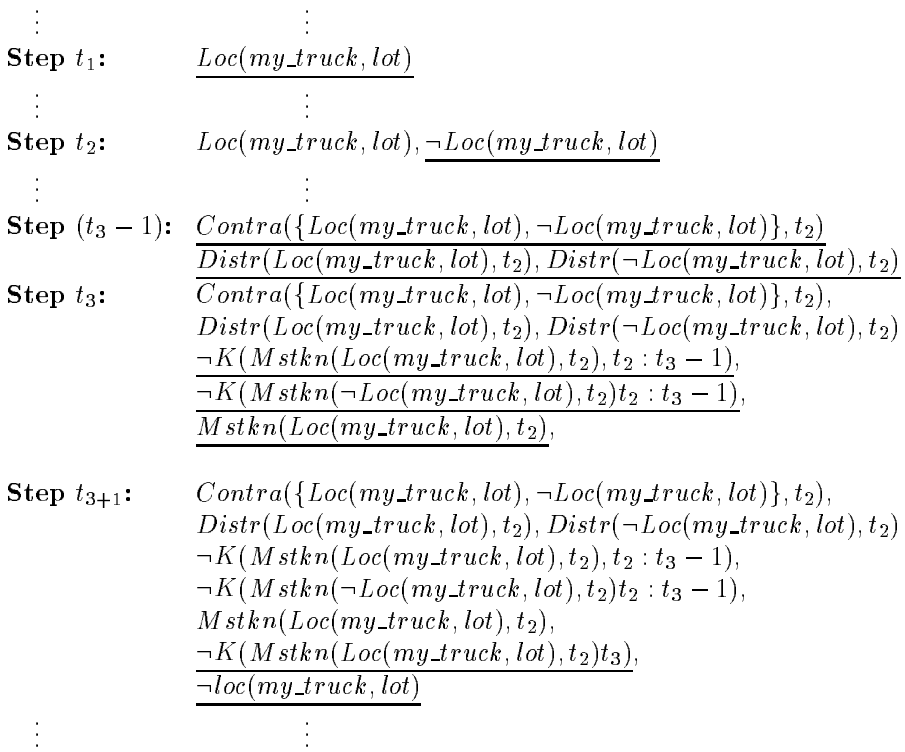

Figure 4.10: A contradictory observation, $\neg loc(my\_truck, lot)$, is properly reinstated.

4.4 Chapter Summary

In this chapter we have introduced the step-logic formalism and added general capabilities to the logics thus far developed. The tools developed here will be used to help solve the *Mistaken Car* and both the *One-* and *Two Johns* problems in the next chapter.

Chapter 5

Problem Solutions

In chapter 3 we introduced and sketched informal solutions to our canonical examples of object-identification errors of the compression type: the *Mistaken Car*, the *One John*, and the *Two Johns* problems. In this chapter we present formal solutions to all three problems using the step-logic tools developed in the previous chapter (chapter 4).

5.1 Notation

In this chapter we will make use of the following notation and abbreviations introduced in the previous chapters (chapters 3 and 4).

- $tfitb(y, S, i)$, i.e., "the thing (object of presentation) which was at first identified to be ..." abbreviates $\iota(x)FITB(x, y, S, i)$
- $MISID(t, S, i)$ abbreviates $tfitb(t, S, i) \neq t$
- $trta(y, i)$, i.e., "the unique thing referred to as y prior to step i" abbreviates $\iota x RTA(x, y, i)$
- **AM:** $(\forall x)(\exists yzij)[(Names(x, y, i) \ \wedge \ Names(x, z, j) \ \wedge y \neq z \ \wedge \ i \leq j) \rightarrow Amb(x, j)]$
- α abbreviates $\alpha[\emptyset]$
- $\vdash_i \alpha < S_1, \ldots, S_n >$ if and only if $\vdash_i \alpha[S_1], \ldots, \alpha[S_n]$ and there is no S such that $\forall k$, $1 \leq k \leq n$, $S \neq S_k$ and $\vdash_i \alpha[S]$

Additionally $\alpha_{(t/t_1)}$ will be used to denote the wff resulting after substituting t_1 for every occurrence of t in α.[1]

5.2 The Formalism for *Mistaken Car*

To reiterate the *Mistaken Car* problem: An agent walks up to what he thinks is his car in the parking lot, and then later a friend tells him that he is mistaken: it's not his car after all.

The inference function we will use, called Obs_{mc}, is defined by:

$$Obs_{mc}(j) = \begin{cases} C(mc, bl), F(mc, mk), \forall xy[At(x, y) \rightarrow P(x, y)] & \text{if } j = 1 \\ At(mc, mk) & \text{if } j = 2 \\ \neg F(mc, mk) & \text{if } j = 3 \\ MISID(mc, \{At(mc, l), \neg F(mc, mk)\}, 2) & \text{if } j = 5 \\ \emptyset & \text{otherwise} \end{cases}$$

where C, F, and bl can be thought of as abbreviations for $Color$, $Fits$, and $blue$ (as in the discussion of chapter 3). $\forall xy[At(x, y) \rightarrow P(x, y)]$ is used here instead of the slightly more complicated wff which appeared earlier, namely $\forall xyz[At(x, y) \rightarrow (y \neq z \ \vee \ \neg AT(X, z))]$. The idea is the same though; to let misidentifications cause error to be propagated through extended MP, in this case to the consequentially mistaken theorem $P(mc, l)$. (For the sake of brevity we have omitted the beliefs $Make(mc, toyota)$ and $Registered(mc, maryland)$ which also appeared in the text of chapter 3. The omitted beliefs would be treated like $C(mc, bl)$ in what follows.)

The step numbers for the introduction of the these observations have been chosen to correspond roughly to those in the earlier text. Most importantly the tutorial

$$MISID(mc, \{At(mc, l), \neg F(mc, mk)\}, 2))$$

is introduced after the presentation of the mistaken car, which occurs at step 2 (when the car is first observed by the reasoner), and after the beliefs produced by the presentation (i.e., $At(mc, mk)$ and $\neg F(mc, mk)$) have been observed. Step 5, at which the tutorial is introduced, is after a direct contradiction occurs (specifically the contradictands $F(mc, mk)$

[1] Renaming variables which appear in t_1 when necessary. Reality terms, which contain ι-variables, are subject to this possibility.

and $\neg F(mc, mk)$ will both appear at step 4) simulating a reasoner who is at first baffled and then comes to correct his mistake (with the aid of the tutor's advice).

The inference function we will use, Inf_{Misid}, contains the rules shown in figure 5.1.[2] Rules 1-6 have already been discussed, though we have not yet seen the annotated version of rule 3 (extended MP). The derivation of this rule is analogous to that of MP which we have seen already; all antecedent theorems together together their derivations comprise the derivation of the consequent of both rules. (See note (*) in the footnote of figure 5.1.) Also note that additional stipulations placed on INH and extended MP (notes a,b) now prevent the applicability of these rules to $Mstkn$ beliefs as well as $Distr$-ed beliefs.

Rules 7 and M are both new.

Rule 7: This rule is analogous to rule 5 of Inf_{deriv} which tracks down the consequences of $Distr$-ed beliefs and marks those consequences as $Distr$-ed. Rule 7 here tracks down the consequences of $Mstkn$ beliefs and marks them as $Mstkn$. There is no need for a rule to track down the antecedents of a $Mstkn$ belief, similar to rule 6 of Inf_{deriv}, since all mistakes are assumed to stem initially from a $MISID$-ed observation. Thus the agent will always come to know that an antecedent belief is $Mstkn$ before she comes to know that a consequence of that belief is $Mstkn$.

Rule M: This rule takes care of the renaming of a misidentified object in the beliefs produced by the presentation. It says this: If α, containing the term t, was produced by a presentation at step k and a misidentification of t comes to the reasoner's attention at a later step i, then at $i+1$ the reasoner will believe that α holds of the misidentified object (of presentation), i.e., $tfitb(t, S, k)$ where $\alpha \in S$. The rule also marks the offending belief(s) $Mstkn$ as of the step of presentation.

5.2.1 Solution to *Mistaken Car*

The solution is given in figure 5.2.1. The rule used for each newly inferred theorem is given just to the left of that theorem (except for those inherited from one step to the next via rule 2, i.e., these are the non-underlined theorems). Only those theorems of interest are shown; some irrel-

[2] Not shown in this figure are rules 5 and 6 from Inf_{deriv}. They are not needed to solve the *Mistaken Car* problem defined by Obs_{mc} since the contradictands in this problem are both observations, neither of which lead to or were caused by other beliefs (via MP). In more general settings these additional rules may be needed, and in such cases should be included.

Rule 1:	$\mathbf{i}$: $\quad$ $\mathbf{i+1}$: α	If $\alpha \in OBS(i+1)$
Rule 2:	$\mathbf{i}$: $\alpha[S]$ $\mathbf{i+1}$: $\alpha[S]$	Inheritance[a]
Rule 3:	$\mathbf{i}$: $\alpha_1(t)[S_1], \ldots, \alpha_n(t)[S_n], (\forall x)[\alpha_1(x) \wedge \ldots \wedge \alpha_n(x)] \rightarrow \beta(x)[S_{n+1}]$ $\mathbf{i+1}$: $\beta(t)[S]^*$	Extended MP[b]
Rule 4:	$\mathbf{i}$: $\alpha[S_1],\ \neg\alpha[S_2]$ $\mathbf{i+1}$: $Contra(\alpha, \neg\alpha, i), Distr(\alpha, i), Distr(\neg\alpha, i)$	Contra & Distr Noted
Rule 5:	$\mathbf{i}$: $Distr(\alpha, j)$ $\mathbf{i+1}$: $\neg K(Mstkn(\alpha, j), i)$	Negative Introspection[c]
Rule 6:	$\mathbf{i}$: $Contra(\{\alpha, \beta\}, k), Mstkn(\beta, j), \neg K(Mstkn(\alpha, k), i-1)$ $\mathbf{i+1}$: α	Reinstatement
Rule 7:	$\mathbf{i}$: $\alpha < S_1, \ldots, S_m >, Mstkn(\beta_1, j), \ldots, Mstkn(\beta_n, j)$ $\mathbf{i+1}$: $Mstkn(\alpha, i)$	Mistaken Consequences[d]
Rule M:	$\mathbf{i}$: $MISID(t, S, k)$ $\mathbf{i+1}$: $\alpha_{(t/tfitb(t,S,k))}, Mstkn(\alpha, k)$	Misid Renaming[e]

* Where S is $\{\alpha_1(t), \ldots, \alpha_n(t), (\forall x)[(\alpha_1(x) \wedge \ldots \wedge \alpha_n(x)) \rightarrow \beta(x)]\} \cup S_1 \cup \ldots \cup S_{n+1}$

[a] Where $\not\vdash_i Distr(\alpha, i-1)$,$\not\vdash_i Mstkn(\alpha, i-1)$ $\not\vdash_i \neg\alpha$, and for each $\beta \in S$ $\not\vdash_i Distr(\beta, i-1)$ and $\not\vdash_i Mstkn(\alpha, i-1)$. Also, if α is of the form $\neg\gamma$ then this rule does not apply if $\vdash_i \gamma$.

[b] The stipulations placed on the antecedent, $\alpha[S]$, of INH in note (a) above apply to each antecedent wff here, as well as to $\beta(t)$.

[c] Where $\not\vdash_i Mstkn(\alpha, j)$.

[d] Where each S_k contains at least one of $\beta_1, \ldots \beta_n$.

[e] Where $\alpha \in S$.

Figure 5.1: Inf_{Misid} – for correcting compression-based identification errors.

evant wffs involving negative introspection are omitted. The theorem $\forall xy[At(x,y) \rightarrow P(x,y)]$, introduced as a 1-observation, is abbreviated by A in the figure for brevity. Derivations are set in bold type. The comments to the right paraphrase the reasoning process.

5.3 The Formalism for *One John*

Recall the *One John* scenario from chapter 3: we imagine that the agent is talking to Sally about a third person, whom the agent initially comes to identify as his (unmarried) friend John, merely in virtue of matching John to Sally's description of the person, or the context of the conversation, etc., but not in virtue of hearing Sally use the name "John". Later he finds out it is not John, but someone else.

Proper naming and the use of names is made explicit with the the predicate symbol $Names$ (discussed in chapter 3).

The inference function we will use, Obs_{1j}, is defined by:

$$Obs_{1j}(i) = \begin{cases} \neg M(j), Names(`j, j, -\infty) & \text{if } i = 1 \\ M(j), B(j) & \text{if } i = 2 \\ MISID(mc, \{M(j), B(j)\}, 2) & \text{if } i = 4 \\ \emptyset & \text{otherwise} \end{cases}$$

where M, B, and j can be thought of as abbreviations for $Married$, $BrokenLeg$, and $john$, respectively. $Names(`j, j, -\infty)$ is used to indicate that $j$ first came to be known as $`j$ by the agent at some arbitrarily distant time in the past (or, that the agent has "always" known j by the name $`j$).

5.3.1 Solution to *One John*

The reasoning proceeds much like in the *Mistaken Car*. It is illustrated in the figure (5.3). In the figure ellipses (. . .) appearing at step $i+1$ indicate that *all* i-theorems are inherited, and hence reappear, as $i+1$-theorems.)

Step 0: $\emptyset$

Step 1:
(R1) $\underline{C(mc, bl), F(mc, mk), A}$ **Agent:** "My car is a blue ..."

Step 2: $C(mc, bl), F(mc, mk), A,$
(R1) $\underline{At(mc, l)}$ **Agent:** "Oh, there's my car!"

Step 3: $C(mc, bl), F(mc, mk), A, At(mc, l),$
(R3) $\underline{P(mc, l)}$ $[\{\mathbf{At(mc, l)}, \mathbf{A}\}],$
(R1) $\underline{\neg F(mc, mk)}$ **Agent:** "Hey, my key doesn't fit."

Step 4: $C(mc, bl), A, At(mc, l),$
$P(mc, l)$ $[\{\mathbf{At(mc, l)}, \mathbf{A}\}],$
(R4) $\underline{Contra(\{F(mc, mk), \neg F(mc, mk)\}, 3),}$ **Agent:** "That's impossible!"
(R4) $\underline{Distr(F(mc, mk), 3), Distr(\neg F(mc, mk), 3)}$

Step 5: $C(mc, bl), A, At(mc, l),$
$P(mc, l)$ $[\{\mathbf{At(mc, l)}, \mathbf{At(mc, l)} \rightarrow \mathbf{P(mc)}\}],$
$Contra(\{F(mc, mk), \neg F(mc, mk)\}, 3),$
$Distr(F(mc, mk), 3), Distr(\neg F(mc, mk), 3),$
(R1) $\underline{MISID(mc, \{At(mc, l), \neg F(mc, mk)\}, 2)}$ **Friend:** "This isn't you're car."

Step 6 $C(mc, bl), A, At(mc, l),$
$P(mc, l)$ $[\{\mathbf{At(mc, l)}, \mathbf{A}\}],$
$Contra(\{F(mc, mk), \neg F(mc, mk)\}, 3),$
$Distr(F(mc, mk), 3), Distr(\neg F(mc, mk), 3),$
$MISID(mc, \{At(mc, l), \neg F(mc, mk)\}, 2),$
(RM) $\underline{Mstkn(At(mc, l), 2), Mstkn(\neg F(mc, mk), 2),}$
(RM) $\underline{At(tfitb(mc, \{At(mc, l), \neg F(mc, mk)\}, 2),}$ **Agent:** "I mistook this car for mine."
(RM) $\underline{\neg F(tfitb(mc, \{At(mc, l), \neg F(mc, mk)\}, 2),}$
(R5) $\underline{\neg K(Mstkn(F(mc, mk), 3), 5)}$ **Agent:** "Hmmm, let me see..."

Step 7 $C(mc, bl), At(mc, l) \rightarrow P(mc), At(mc, l),$
$Contra(\{F(mc, mk), \neg F(mc, mk)\}, 3),$
$Distr(F(mc, mk), 3), Distr(\neg F(mc, mk), 3),$
$MISID(mc, \{At(mc, l), \neg F(mc, mk)\}, 2),$
$Mstkn(At(mc, l), 2), Mstkn(\neg F(mc, mk), 2),$
$Mstkn(P(mc), 2), At(tfitb(mc, \{At(mc, l), \neg F(mc, mk)\}, 2),$
$\neg F(tfitb(mc, \{At(mc, l), \neg F(mc, mk)\}, 2),$
$\neg K(Mstkn(F(mc, mk), 3), 5),$
(R6) $\underline{F(mc, mk)}$ **Agent:** "Then my key is ok and..."
(R7) $\underline{Mstkn(P(mc, l), 2),}$
(R3) $\underline{P(tfitb(mc, \{At(mc, l), \neg F(mc, mk)\}, 2), l)}$
$[\{\mathbf{At(tfitb(mc, \{At(mc, l), \neg F(mc, mk)\}, 2), l), A}\}]$

Figure 5.2: Solution to the *Mistaken Car* problem.

Step 0: $\emptyset$

Step 1:

(R1) $\underline{\neg M(j), Names(`j, j, -\infty)}$

Step 2:

(R1) $\ldots \underline{B(j), M(j)}$ **Sally:** "His leg is broken and his wife ..."

Step 3:

(R4) $Names(`j, j, -\infty), \underline{Contra(\{M(j), \neg M(j)\}, 2)},$

(R4) $\underline{Distr(M(j), 2), Distr(\neg M(j), 2)}$ **Agent:** "That's impossible!"

Step 4:

(R1) $\ldots \underline{MISID(j, \{M(j), B(j)\}, 2)}$ **Sally:** "You've misidentified who I'm talking about."

Step 5:

(RM) $\ldots \underline{Mstkn(M(j), 2), Mstkn(B(j), 2)},$

(RM) $\underline{M(tfitb(j, \{M(j), B(j)\}, 2), B(tfitb(j, \{M(j), B(j)\}, 2)},$

(R5) $\underline{\neg K(Mstkn(M(j), 3), 4)}$ **Agent:** "So that's what's wrong."

Step 6:

(R6) $\ldots \underline{\neg M(j)}$ <Reinstate Marital Belief>"

Figure 5.3: Solution to the *One John* problem.

5.4 The Formalism for *Two Johns*

We now suppose that subsequent to step 6 (from *One John*, above) Sally tells the agent that she is talking about someone named "John" (different from the agent's friend John). The agent thus becomes aware that "John" is ambiguous, and again must do some renaming.

To make the role of $Names$ more precise inference rule N (below) is used.[3]

Rule N: $\mathbf{i}: \dfrac{Names(`x, y, k), \alpha(x)}{\mathbf{i+1}: \ \alpha_{(x/y)}}$ USE OF NAMES

Rule N is grounded in the interplay between the *use* and *mention* of standard names. It makes explicit that the function of a name when used is to stand for an object. The rule says this: If the agent believes that α of some object x (a *use* of x) and she believes that the standard name of that object (the quoted form, $`x$) names some object y, possibly different from the first (a *mention* of x), then the agent will come to believe that α of y. For instance, since "Merle" names my sister (I believe that) then if I come to believe that Merle is in town, I'll also come to believe that my sister is in town.

When a name, $`x$, does not uniquely refer, use of that name is confusing. This too is reflected in rule N. For instance, ignoring the third argument of $Names$ for the moment, if $Names(`j, smith)$, $Names(`j, jones)$, and $smith \neq jones$ are all i-theorems, then the non-judicious use of rule N will produce both $jones \neq jones$ and $smith \neq smith$ as $i+1$-theorems. A cautious reasoner will refrain from *using* the ambiguous term (at least in ambiguous contexts) thereby circumventing this difficulty.

Recall that the predicate symbol **Amb** is used to state that a name does not refer uniquely beyond a certain step. The renaming on ambiguous terms is accomplished via rule A below:

Rule A: $\mathbf{i}: \dfrac{Amb(`x, k), \alpha(x)}{\mathbf{i+1}: \ \alpha_{(x/trta(`x,k))}}$ AMBIGUITY RENAMING

[3] Like the rules of MP and INH, the proper conditions for the application N will need to be stipulated.

This rule takes an antecedent wff $\alpha(x)$ which *uses* the ambiguous term x and eliminates the offending term replacing it with $trta(`x, k)$, which mentions but does not use x.

The inference function, $Inf_{Disambig}$, used to solve *Two Johns* contains those of Inf_{Misid} (with the appropriate modifications to the stipulations to limit the use of ambiguous terms) plus rules N and A (see figure 5.4).[4]

The new observation function, Obs_{2j}, adds two observations, namely **AM** and

$$Names(`j, tfitb(j, \{M(j), B(j)\}, 2), 2)$$

to those appearing in Obs_{1j}. $Names(`j, tfitb(j, \{M(j), B(j)\}, 2), 2)$ says of the object the agent took to be 'his' John, that it is named "John". The addition of this theorem introduces the ambiguity. $Obs_{2j}$ is shown below. We pick step 7, the first step beyond the resolution of *One John*, for the introduction of $Names(`j, tfitb(j, \{M(j), B(j)\}, 2), 2)$.[5] Any later step would also be fine.

$$Obs_{2j}(i) = \begin{cases} \neg M(j), Names(`j, j, -\infty), AM & \text{if } i = 1 \\ M(j), B(j) & \text{if } i = 2 \\ MISID(j, \{M(j), B(j)\}, 2) & \text{if } i = 4 \\ Names(`j, tfitb(j, \{M(j), B(j)\}, 2), 2) & \text{if } i = 7 \\ \emptyset & \text{otherwise} \end{cases}$$

5.4.1 Solution to *Two Johns*

The solution to *Two Johns* is depicted in figure 5.5. We pick up from step 6, the last step in the solution to *One John* (shown in figure 5.3) adding **AM** as a 1- through 6-theorem, as per Obs_{2j}. For the sake of brevity we have omitted the derivation of $Amb(`j, 2)$ which is:

[AM, Names('j,j), Names('j,tfitb(j,{M(j), B(j)},2)), MISID(j, {M(j),B(j)},2), $-\infty \leq 2$]

None of the omitted derivations have any effect on the reasoning. The formal treatment of the theorem $-\infty \leq 2$, found in the derivation above (it appears because it binds to the last conjunct of the antecedent of **AM**) has been omitted as well.

---

[4] As well as extended modus ponens which is required to infer $Amb(`j, 2)$ from the axiom **AM**.

[5] This theorem must necessarily be introduced after the agent comes to know about the second John (step 4), but not necessarily before he has resolved his initial misidentification, as is done here. The current formalism, however, requires this additional restriction.

Rule 1:	$\mathbf{i}:$ $\_\_\_\_\_$ $\mathbf{i+1}:$ α		IF $\alpha \in OBS(i+1)$
Rule 2:	$\mathbf{i}:$ $\alpha[S]$ $\mathbf{i+1}:$ $\alpha[S]$		INHERITANCE[a,d]
Rule 3:	$\mathbf{i}:$ $\alpha_1(t)[S_1], \ldots, \alpha_n(t)[S_n], \forall x[(\alpha_1(x) \wedge \ldots \wedge \alpha_n(x)) \rightarrow \beta(x)][S_{n+1}]$ $\mathbf{i+1}:$ $\beta(t)[S]^*$		EXTENDED MP[b,d]
Rule 4:	$\mathbf{i}:$ $\alpha[S_1],\ \neg\alpha[S_2]$ $\mathbf{i+1}:$ $Contra(\alpha, \neg\alpha, i), Distr(\alpha, i), Distr(\neg\alpha, i)$		CONTRA & DISTR NOTED
Rule 5:	$\mathbf{i}:$ $Distr(\alpha, j)$ $\mathbf{i+1}:$ $\neg K(Mstkn(\alpha, j)i)$		NEGATIVE INTROSPECTION[c]
Rule 6:	$\mathbf{i}:$ $Contra(\{\alpha, \beta\}, k), Mstkn(\beta, j), \neg K(Mstkn(\alpha, k), i-1)$ $\mathbf{i+1}:$ α		REINSTATEMENT[d]
Rule 7:	$\mathbf{i}:$ $\alpha < S_1, \ldots, S_m >, Mstkn(\beta_1, j), \ldots, Mstkn(\beta_n, j)$ $\mathbf{i+1}:$ $Mstkn(\alpha, i)$		MISTAKEN CONSEQUENCES[e]
Rule M:	$\mathbf{i}:$ $MISID(t, S, k)$ $\mathbf{i+1}:$ $\alpha_{(t/tfitb(t,S,k))}, Mstkn(\alpha, k)$		MISID RENAMING[f]
Rule N:	$\mathbf{i}:$ $Names(`x, y, k), \alpha(x)$ $\mathbf{i+1}:$ $\alpha_{(x/y)}$		USE OF NAMES[g]
Rule A:	$\mathbf{i}:$ $Amb(`x, k), \alpha(x)$ $\mathbf{i+1}:$ $\alpha_{(x/trta(`x,k))}$		AMBIGUITY RENAMING

* Where S is $\{\alpha_1(t), \ldots, \alpha_n(t), \forall x[(\alpha_1(x) \wedge \ldots \wedge \alpha_n(x)) \rightarrow \beta(x)]\} \cup S_1 \cup \ldots \cup S_{n+1}$

[a] Where $\nvdash_i Distr(\alpha, i-1)$,$\nvdash_i Mstkn(\alpha, i-1)$ $\nvdash_i \neg\alpha$, and for each $\beta \in S$ $\nvdash_i Distr(\beta, i-1)$ and $\nvdash_i Mstkn(\alpha, i-1)$. Also, if α is of the form $\neg\gamma$ then this rule does not apply if $\vdash_i \gamma$.

[b] The stipulations placed on the antecedent, $\alpha[S]$, of INH in note (a) above apply to each antecedent wff here, as well as to $\beta(t)$.

[c] Where $\nvdash_i Mstkn(\alpha, j)$.

[d] Where α does not *use* the term y if $\vdash_i Amb(`y, k)$. (For rule 3 this applies to each α_j and β.)

[e] Where each S_k contains at least one of $\beta_1, \ldots \beta_n$.

[f] Where $\alpha \in S$.

[g] Where $\nvdash_i x \neq y$.

Figure 5.4: $Inf_{Disambig}$ – for disambiguating after a dispersion-based misidentification

$\vdots$ $\vdots$

Step 6:

$Names(`j, j, -\infty), Contra(\{M(j), \neg M(j)\}, 2), AM,$
$Distr(M(j), 2), Distr(\neg M(j), 2),$
$MISID(j, \{M(j), B(j)\}, 2), Mstkn(M(j), 2), Mstkn(B(j), 2),$
$M(tfitb(j, \{M(j), B(j)\}, 2), B(tfitb(j, \{M(j), B(j)\}, 2),$
$\neg K(Mstkn(M(j), 2), 4), \underline{\neg M(j)}$

Step 7:

(R1) $\ldots \underline{Names(`j, tfitb(j, \{M(j), B(j)\}, 2), 2)}$ **Sally:** "His name is 'John"'

Step 8:

(R3) $\ldots \underline{Amb(`j, 2)}$ **Agent:** "Oh, they have the same name.

Step 9:

$Contra(\{M(j), \neg M(j)\}, 2), AM, Distr(M(j), 2), Distr(\neg M(j), 2),$
$Mstkn(M(j), 2), Mstkn(B(j), 2), \neg K(Mstkn(M(j), 2), 4), Amb(`j, 2),$
(RA) $\underline{Names(`j, tfitb(trta(`j, 2), \{M(j), B(j)\}, 2), 2), Names(`j, trta(`j, 2), -\infty)},$
(RA) $\underline{\neg M(trta(`j, 2)), MISID(trta(`j, 2), \{M(j), B(j)\}, 2)},$
(RA) $\underline{M(tfitb(trta(`j, 2), \{M(j), B(j)\}, 2), B(tfitb(trta(`j, 2), \{M(j), B(j)\}, 2)}$

Figure 5.5: Solution to the *Two Johns* problem – continued from the solution to *One John*.

Chapter 6

Indexicality: A Case Study for FOL

We have now completed our discussion of our terminological change work and turn our attention to a separate discussion of indexicality. An indexical is an expression whose referent is dependent on the context in which that expression is used; it's meaning changing as the context changes. "I", "now", and "here" are examples. We have already touched on the changing meaning of "now" with our discussion of step-logics. We have seen also that even proper names can have an indexical nature with our discussion of presentations and *this* and *that* (chapter 3). Our discussion now turns to the indexicality of the "I".

Though this is a separate theme from that which we just covered we are still concerned with reasoned change in belief, but here the change is not so much over time or proper names, but rather over context of who is uttering a phrase. Even though this treatment uses classical first-order logic rather than step-logics, it shares a feature with the preceding proper names treatment in that both treatments show that dealing with indexical or changing meanings is far trickier than one might expect.

Our treatment of "I" here is based on a paper by Ohlbach in which he shows that a FOL representation of knowledge about indexicals is non-trivial. He, and we in turn, discuss the indexical "I" in the context of one of Smullyan's logic puzzles which illustrates sharply how the meaning of "I" changes with context.

6.1 Knights and Knaves

The *Knights and Knaves* problem ([Smullyan, 1978]) can be stated as follows: An island exists whose only inhabitants are knights, knaves, and a princess. The knights on the island always tell the truth, while the knaves always lie. Some of the knights are poor and the rest of them are rich. The same holds for the knaves. The princess is looking for a husband who must be a rich knave. In uttering one statement, how can a rich knave convince the princess that he is indeed a prospective husband for her? [1]

In [Ohlbach, 1984] we find the framing and solution of this problem in a formal theorem-proving context using first-order logic (FOL). Though trying to write the problem in FOL may not appear to be difficult at first, it is shown by Ohlbach not to be entirely elementary. He examines, and finds inadequate, two different approaches before he finally settles on a third. This final approach, though successful in that it gets the desired "solution", is unsatisfactory. Specifically, Ohlbach uses a truth predicate with two arguments, $T(x, y)$, which asserts that its first argument x is true, but has no clear meaning for its second argument in terms of the original problem. Its justification is that the predicate allows a theorem-prover to perform certain unifications that lead to the intended solution. But it does not accomplish the goal of finding a knowledge representation faithful to the original problem, as well as having the solution as a logical consequence.

Ohlbach's conclusion is that knowledge representation is too hard in FOL, and too dependent upon tricks. We will not dispute that it takes some time to come up with a satisfactory representation of the problem, but this is not necessarily the fault of FOL. We contend that there is a straightforward treatment of the problem that is faithful to its intent and that does allow a formal proof of the desired result. However, it requires employing concepts that are not usually found in the context of problem-solving via resolution theorem-provers, namely, ideas from natural language processing. Nevertheless, we are not replacing one trick by another, but rather introducing a well-understood and general formalism for problems of this sort.

The rest of this chapter is organized as follows: Section 6.2 discusses issues of problem rep-

[1] The intended solution is the statement "I am a poor knave." The reader can readily verify that this indeed is a solution. Note the self-referential nature of the statement; this feature is a special case of indexicality, which we address below. For general treatments of self-reference, see [Perlis, 1985], where another of Smullyan's puzzles is treated, and [Smith, 1986].

resentation, especially the role played by the pronoun "I" in the *Knights and Knaves* problem. Section 6.3 reviews general consequences for truth-values of statements containing indexicals such as "I", and section 6.4 applies this to the *Knights and Knaves* problem. Section 6.5 gives our formal treatment, including a resolution-refutation proof by answer extraction. Section 6.6 compares our solution to Ohlbach's and suggests a broader context for dealing with self-utterances in automatic theorem-proving.

6.2 Problem Representation

Finding a suitable representation for problems in AI is often a difficult task. However, the formalism used to represent a problem is not necessarily the cause of the difficulty, though we grant that sometimes it is. Often it is the problem itself that is resisting representation and, when this occurs, further insight into the problem is necessary. The *Knights and Knaves* problem is a prime example of this. Ohlbach's interpretation of the problem results in his asking "Is there a statement x that I (being a rich knave) can say to convince the princess that I am indeed a rich knave?" Formally this might look like (and does in Ohlbach's second treatment):[2]

$$\textbf{OHL: } \exists u[CanSay(I,u) \leftrightarrow T(and(knave(I), rich(I)))]$$

where T is the predicate meaning True and "and" although a function symbol, intuitively takes two statements as arguments and returns another single conjunctive statement.[3]

This interpretation may appear to be reasonable given the English statement of the problem, but as Ohlbach discusses, this representation (along with other associated axioms) is not sufficient to derive the intended result.

Part of the difficulty is not hard to see. The constant "I" stands for a fixed person (who is a rich knave). The point of the biconditional in **OHL**, and especially of the right-hand side, is to test whether the speaker is a rich knave, based on the ability to utter u. That is, the

[2] Ohlbach's first treatment involved the axiom $\exists u[CanSay(I,u) \rightarrow T(and(knave(I), rich(I)))]$ which (in addition to yielding a trivial and unhelpful answer) does not seem to correspond to his English interpretation "There exists a statement which I can say and which implies that I am really a rich knave." In fact, it seems to us that the goal statement $\exists u[CanSay(I,u) \wedge \forall p[CanSay(p,u) \rightarrow T(and(rich(p), knave(p)))]$ comes much closer to the English.

[3] Actually, a name of the statement.

problem really seems to be asking, "What statement, when made by *anyone*, will convince the princess that *the person making the statement* is a rich knave?" The first problem with **OHL** is then the following: "I" should not be bound to a fixed individual, but should represent any "man in the street" who might utter u. We suggest the alternative version:

$$\textbf{G: } \exists u \forall p[CanSay(p, u) \leftrightarrow T(and(rich(p), knave(p)))]$$

We claim to have now adequately represented the goal statement;[4] but this is still not enough. For although this goal statement expresses what we want, there are other problems arising from the truth conditions of utterances containing the pronoun "I". These will enter into axioms in the problem representation, rather than the goal statement.

6.3 Utterance Instances of Statements

This brings us to what we think is the key issue in this puzzle, an issue which has broader significance as well. Specifically, utterances are instances of statement uses, and these instances, in general, have truth-values, rather than the statement in and of itself. In particular, terms in a statement may have no definite reference outside the context of an utterance. Although this concept is familiar to linguists[5] and philosophers (it is the so-called problem of indexicals which is discussed below) it is worth going into detail in here, since the issue of representing knowledge in the *Knights and Knaves* problem hinges on this very phenomenon.

Typically, we think of a statement as being either true or false. This, however, is not always the case. For example, the statement: "I am a knave" will have a truth-value dependent upon who the speaker is; and so would be falsely uttered by any knight and truly by any knave.[6] Statements that contain indexicals have meanings, and hence truth-values, that depend upon context. Another example is the statement: "It is raining". In certain cases when uttering this statement the speaker intends to express that it is raining at some particular place at some particular time. The time and place implied in these cases are "here" and "now" respectively;

[4] Both **G** and the second wff in footnote 2 will do equally well.

[5] Including those who work in natural language processing; see for instance [Allen, 1984], [Allen and Perrault, 1980], and [Harper and Charniak, 1986].

[6] Hence, this statement can be uttered by neither knights nor knaves, in the *Knights and Knaves* problem!

"here" and "now" are indexicals whose referents are significant in determining the truth or falsity of such situated statements.

Generally speaking, an indexical in an utterance is a sub-expression of that utterance whose meaning is determined (and thus understood) by the context in which the utterance is stated. Because of the indeterminacy of truth-values of sentences that contain indexicals, we will refer only to the truth-value of utterance-instances of such statements. An utterance-instance of a statement contains a context in which the statement was (or is) made including who the utterer is.

6.4 "Who Am I?"

If we look closely at any of Ohlbach's representations of the *Knights and Knaves* problem, we notice that the constant "I" seems to be playing two different roles. In all of his goal statements "I" is presumably used as the name of a particular person. Ohlbach's second goal statement, **OHL**, illustrates this usage. On the other hand, in the intended solution to the problem, the "u" of the goal statement is bound to the term:

$$and(not(rich(I)), knave(I))$$

(which we abbreviate by *anriki*) where, the same symbol "I" appears as before but now what is of interest is its potential presence within $CanSay(I, anriki)$, i.e., as part of a potential utterance whose truth value depends upon who the speaker is. That is, any number of people might utter *anriki*, and its meaning would be different in each case. We now have an utterance-instance and need to know who "I" is before assigning a truth-value. "I" must be viewed as a pronoun and not a proper name here. In particular, the knighthood or "knavehood" of "I" determines the truth of *anriki*. Of course, in the world in question, only knaves (and rich ones at that) could utter *anriki*. But that is the point; the princess must be able to deduce precisely that fact: that anyone at all who utters *anriki* must consequently be a rich knave.

In what follows, we remove this ambiguity by introducing a new binary predicate letter, **TU**, into the language of *Knights and Knaves*. *TU*'s first argument is intended to denote a person and its second an utterance. Intuitively $TU(p, u)$ is true if and only if u is true when uttered by person p. In particular, when u contains "I" as a subexpression, $TU(p, u)$ is true if

and only if the substitution-instance of u resulting from replacing all occurrences of "I" in u by p is true. Thus the statement:

TU(John,"I am six feet tall")

is true if and only if John (the utterer) is indeed six feet tall.[7]

6.5 Formalization

We now introduce our notation for representing the problem. We use a first-order theory, embellished with a quotation mechanism (see, for example, [Perlis, 1985]) which contains the following terms, predicate letters and function letters:

I : constant (the word "I")

knave : unary function letter (knave(x) stands for the term "x is a knave")

rich : unary function letter

not : unary function letter

and : binary function letter

CanSay : binary predicate letter ($CanSay(p, u)$ means "p can say u")

TU : binary predicate expression ($TU(p, u)$ means term "u" would be true if occurrences of "I" in u are replaced by p)

T : unary predicate expression ($T(t)$ means the term t is true)

Using the above notation, we can now present the axioms which capture the *Knights and Knaves* problem as we see it. For simplicity, we suppose all variables range over knights, knaves, the princess, and utterances.[8]

[7] This is somewhat comparable to the formulation of Barwise and Perry [1983] when they speak of an utterance in a "situation" concerning "I": u[I am six feet tall]e is true (where u is the utterance "I am six feet tall" and e is a situation in which John is present and makes utterance u) iff John is indeed six feet tall in situation e.

We use TU as an acronym for "truly utters"; i.e., TU(p,u) says "p would be telling the truth if p were to utter u."

[8] This follows the convention of Ohlbach. The use of either typed or relativized variables would eliminate unusual readings at the expense of more complex formulae.

All clauses required to solve this problem can be derived from only three first-order axioms and one schema which are sufficient to represent the needed facts about the world in which the knights, knaves, and princess live, namely:

(1) $\forall pu[T(knave(p)) \leftrightarrow (CanSay(p,u) \leftrightarrow \neg TU(p,u))]$

i.e., p is a knave iff the things u that p can say are precisely those which would be false if p uttered them.

(2) $\forall yz[T(and(y,z)) \leftrightarrow (T(y) \wedge T(z))]$

This captures the meaning of the function letter *and*.

(3) $\forall s[T(s) \leftrightarrow \neg T(not(s))]$

This axiom captures the meaning of the function letter *not*.

(4) $\forall p[TU(p, f(I)) \leftrightarrow T(f(p))]$

Here f is a function variable. An instance of this schema with *rich* replacing f is $TU(Bill, rich(I)) \leftrightarrow T(rich(Bill))$.[9]

Since (4), above, is a functional axiom schema, (ordinary) theorem provers would have to be given a mechanism to select substitution instances in some appropriate fashion. In order to avoid this added difficulty (although it should not be computationally very expensive in this case) we will continue our analysis in terms of a finite axiomatization of this schema, which requires no such mechanism. The following four axioms recursively establish all possible instances of schema (4) in terms of the leftmost function symbol occurring in TU's second argument.

TU_{and}: $\forall uvp[TU(p, and(u,v)) \leftrightarrow (TU(p,u) \wedge TU(p,v))]$

TU_{not}: $\forall up[TU(p, not(u)) \leftrightarrow \neg TU(p,u)]$

TU_{rich}: $\forall p[TU(p, rich(I)) \leftrightarrow T(rich(p))]$

TU_{knave}: $\forall p[TU(p, knave(I)) \leftrightarrow T(knave(p))]$

Figure 6.1 shows axioms (1)-(3) and the above four TU axioms rewritten in clause form to be used in our resolution proof of the solution to *Knights and Knaves*. For ease in reading we omit those clauses which have no bearing on our proof.

[9] Note that this schema is akin to Tarski's convention $T(\alpha) \leftrightarrow \alpha$ in cases where α contains the indexical "I".

KS1: $\neg T(knave(p)) \vee \neg CanSay(p,u) \vee \neg TU(p,u)$
KS2: $\neg T(knave(p)) \vee CanSay(p,u) \vee TU(p,u)$
KS3: $T(knave(p)) \vee \neg CanSay(p,u) \vee TU(p,u)$

A1: $\neg T(and(y,z)) \vee T(y)$
A2: $\neg T(and(y,z)) \vee T(z)$
A3: $T(and(y,z)) \vee \neg T(y) \vee \neg T(z)$

N1: $T(s) \vee T(not(s))$
N2: $\neg T(s) \vee \neg T(not(s))$

TU1: $TU(p,and(u,v)) \vee \neg TU(p,u) \vee \neg TU(p,v)$
TU2: $\neg TU(p,and(u,v)) \vee TU(p,u)$
TU3: $\neg TU(p,and(u,v)) \vee TU(p,v)$
TU4: $TU(p,not(u)) \vee TU(p,u)$
TU5: $\neg TU(p,not(u)) \vee \neg TU(p,u)$
TU6: $TU(p,rich(I)) \vee \neg T(rich(p))$
TU7: $\neg TU(p,rich(I)) \vee T(rich(p))$
TU8: $TU(p,knave(I)) \vee \neg T(knave(p))$
TU9: $\neg TU(p,knave(I))) \vee T(knave(p))$

Figure 6.1: Clause form of axioms for use in solution to the *Knights and Knaves* problem.

We are now ready for the clauses which represent our goal statement. In line with our earlier discussion, we take as our goal statement:

$$\mathbf{G} : \exists u \forall p[CanSay(p, u) \leftrightarrow T(and(rich(p), knave(p)))]$$

In the clauses that follow, "g" is a Skolem function resulting from the elimination of the existential quantifier in the negation of **G**.

G1: $CanSay(g(u), u)) \vee T(and(rich(g(u)), knave(g(u))))$

G2: $\neg CanSay(g(u), u) \vee \neg T(and(rich(g(u)), knave(g(u))))$

Figure 6.2 depicts our resolution proof with answer extraction showing the desired solution, *anriki*, to the *Knights and Knaves* problem. In the figure axioms are abbreviated by their names given above (e.g. **TU1**, etc.). Clauses that are the result of a step in the proof are named (R1, R2, F13, etc.) so that they may referred to later in the proof. For compactness *and* is abbreviated by a, *rich* by r, *knave* by k, *not* by n, and $CanSay$ by CS. Parentheses are eliminated when possible. Key substitutions in factorization steps are shown in boxes beside the resultant clause in which they are instituted. Fac abbreviates "factorization". In clause F13, *anriki*, our desired solution is bound to the Ans term, u. Since this term is not altered in any subsequent steps leading to the derivation of the null clause it has been dropped from clauses R14-R35. It reappears in clause R36 as the solution.

6.6 Discussion

Ohlbach has pointed out an interesting problem in knowledge representation. We agree in principle with his conclusion that knowledge representation is hard. In fact, if someone has to invent a new trick each time they wish to represent a problem, the task would become hopeless. Furthermore, if the language used by the AI practitioner forced the need for tricks, then there would certainly be an argument for selecting another language.

We feel, however, that neither FOL nor automatic theorem-proving imposes any such restriction on the *Knights and Knaves* problem. The complexity that Ohlbach discovered in trying to represent this problem is due to indexicals. In fact, his second argument of the predicate $T(x, y)$ might be dealing with indexical-binding in some way. We have found that

Resolvants	*Resultant Clause*	
A1,		
G1∨Ans(u)	R1:	$CS(gu,u) \vee T(r(gu)) \vee Ans(u)$
N2, R1	R2:	$CS(gu,u) \vee \neg T(nr(gu)) \vee Ans(u)$
A1, R2	R3:	$CS(gu,u) \vee \neg T(a(nr(gu),z)) \vee Ans(u)$
A3, R3	R4:	$CS(gu,u) \vee \neg T(nr(gu)) \vee \neg T(z) \vee Ans(u)$
TU9, R4	R5:	$CS(gu,u) \vee \neg T(nr(gu)) \vee \neg TU(p,k(I)) \vee Ans(u)$
N1, R5	R6:	$CS(gu,u) \vee T(r(gu)) \vee \neg TU(p,k(I)) \vee Ans(u)$
TU6, R6	R7:	$CS(gu,u) \vee TU(gu,r(I)) \vee \neg TU(p,k(I)) \vee Ans(u)$
TU5, R7	R8:	$CS(gu,u) \vee \neg TU(gu,nr(I) \vee \neg TU(p,k(I)) \vee Ans(u)$
TU3, R8	R9:	$CS(gu,u) \vee \neg TU(gu,nr(I) \vee \neg TU(p,a(u',k(I))) \vee Ans(u)$
TU2, R9	R10:	$CS(gu,u) \vee \neg TU(gu,a(nr(I),v)) \vee \neg TU(p,a(u',k(I))) \vee Ans(u)$
R10, *Fac*	F11:	$CS(gu,u) \vee \neg TU(gu,anriki)$ $\boxed{u', v, p \Rightarrow nr(I), k(I), gu}$
KS2, F11	R12:	$CS(gu,u) \vee CS(gu,anriki) \vee \neg T(k(gu)) \vee Ans(u)$
R12, *Fac*	F13:	$CS(g(anriki),anriki) \vee \neg T(k(g(anriki)))$
		$\vee$ **Ans(anriki)** $\boxed{u \Rightarrow anriki}$
G1, A2	R14:	$CS(gu,u) \vee T(k(gu)$
F13, R14	R15:	$CS(g(anriki),anriki)$
G2, R15	R16:	$\neg T(a(r(g(anriki)),k(g(anriki)))$
A3, R16	R17:	$\neg T(r(g(anriki))) \vee \neg T(k(g(anriki))))$
N1, R17	R18:	$T(nr(g(anriki))) \vee \neg T(k(g(anriki)))$
KS3, R18	R19:	$\neg CS(g(anriki),t) \vee TU(g(anriki),t) \vee T(nr(g(anriki)))$
F13, R19	R20:	$TU(g(anriki),anriki) \vee T(nr(g(anriki)))$
TU2, R20	R21:	$TU(g(anriki),nr(I)) \vee T(nr(g(anriki)))$
TU5, R21	R22:	$\neg TU(g(anriki),r(I)) \vee T(nr(g(anriki)))$
TU6, R22	R23:	$\neg T(r(g(anriki))) \vee T(nr(g(anriki)))$
N1, R23	R24:	$T(nr(g(anriki)))$
R15, KS1	R25:	$\neg T(k(g(anriki))) \vee \neg TU(g(anriki),anriki)$
TU1, R25	R26:	$\neg T(k(g(anriki))) \vee \neg TU(g(anriki),k(I)) \vee \neg TU(g(anriki),nr(I))$
TU4, R26	R27:	$\neg T(k(g(anriki))) \vee \neg TU(g(anriki),k(I)) \vee TU(g(anriki),r(I))$
TU7, R27	R28:	$\neg T(k(g(anriki))) \vee \neg TU(g(anriki),k(I)) \vee T(r(g(anriki)))$
N2, R28	R29:	$\neg T(k(g(anriki))) \vee \neg TU(g(anriki),k(I)) \vee \neg T(nr(g(anriki)))$
R24, R29	R30:	$\neg T(k(g(anriki))) \vee \neg TU(g(anriki),k(I))$
TU8, R30	R31:	$\neg T(k(g(anriki)))$
R15, KS3	R32:	$T(k(g(anriki))) \vee TU(g(anriki),anriki)$
R31, R32	R33:	$TU(g(anriki),anriki)$
TU3, R33	R34:	$TU(g(anriki),k(I))$
TU9, R34	R35:	$T(k(g(anriki)))$
R31, R35	R36:	**Ans(anriki)**

Figure 6.2: A resolution solution to the *Knights and Knaves* problem.

a proper treatment of indexical-binding makes for a natural and correct (in that a proper solution is found) representation of the *Knights and Knaves* problem.

Our solution was longer than Ohlbach's. His optimized proof had 20 steps, while ours has 36. Thus the new issues we have introduced into the problem representation have not reduced the complexity; rather they have increased it, but not excessively so. The use of indexicals seems viable within an automatic theorem-proving context.

We feel that this problem is indicative of a whole class of problems that can be handled in a similar fashion, i.e., not dependent upon isolated or ad hoc tricks. In the *Knights and Knaves* problem we defined *TU* in terms of the indexical "I" only. This is because "I" is the only indexical of importance in this problem. In broader contexts, however, this would be insufficient and generalizations of *TU* would be necessary. We offer TU as a step toward a uniform solution to the problem of automatic theorem-proving with indexicals. It will be interesting to see how well generalizations of *TU* handle other indexicals and other problems.

Chapter 7

Typ-constants and Range Defaults

We now turn to the third major theme of this thesis: representing default knowledge. It is widely accepted that default (or non-monotonic) reasoning is endemic in commonsense reasoning. Defaults can be, and often are, viewed as typicality statements of the form "P's are typically Q's". In this chapter we present a different way to view defaults and offer a formal treatment by extending a first-order language to include representations of typical mental notions, in the form of constant symbols called *typicality-* (or simply, *typ-*) constants.

Upon inspection of this formalism we uncover an apparently new aspect of default knowledge, which we call *range* or *irreducible disjunctive* defaults. Here typicality is viewed as spreading over a *range* of possible default conclusions. "Cardinals are typically red or russet" is a reliable default while both "cardinals are typically red" and "cardinals are typically russet" are not. The range "red or russet" is essential, though shown here to require adjustment of previous formalisms.

The difficulty we find is that of representing the denial of default information, for instance that "cardinals are typically red" is *not* a legitimate default. We shall see that representing such denials are important parts of commonsense reasoning but that their representation is not completely straightforward. Although we first observed this for our own *typ*-constant formalism, we show that the phenomenon is quite general to all default formalisms, and yet is unexplored in the literature. We present a formal proposal to solve the range default problem and discuss its shortcomings.

7.1 Typicality

Defaults can be, and often are, viewed as typicality statements of the form "typically X's are Y's". Much effort in AI has been directed toward the development of logic-based approaches for interpreting, representing, and reasoning about typicality statements. A variety of formalisms have been proposed, including, but not limited to, Delgrande's conditional logic[Delgrande, 1987], [Delgrande, 1988] McDermott and Doyle's (NML)[McDermott and Doyle, 1980], Reiter's Default Logic (DL)[Reiter, 1980], and McCarthy's circumscription[McCarthy, 1980]. These, and other, approaches differ markedly from one another in their interpretation of typicality, in technique, and in how successful they are in modeling default reasoning (see [Etherington, 1988] for a good review of the different formalisms). Despite their vast differences, these formalisms share the common view that typicality statements are regarded as *rules of thumb* that are applied, (in a way highly dependent on the formalism), to objects in the "real" domain of the reasoner.

A very different way to view typicality (and hence defaults), and the intuition behind the treatment presented here, is to treat a reasoner's mental concept of a typical or generic instance, which roughly corresponds to a general (indefinite) description, as an object in its own right. That such mental objects have certain properties "encodes" the defaults about the concepts that the objects represent. For example, I have a "mental notion" of what is for me a typical tree. That this typical tree notion (for me) has "leaves" encodes my default that "typically trees have leaves"; that it has "branches" encodes my default that "typically trees have branches". That I rely on the properties of my typical tree notion to (defeasibly) decide about the status of a real tree (i.e., that it too has leaves and branches) encodes the process of jumping to a default conclusion.

In this chapter we formalize this intuition by extending a first-order language to include representations of these mental notions, in the form of constant symbols called *typicality* (or simply, *typ*) constants, which are written as typ_Φ for expressions Φ in the language associated with an indefinite description. As reified objects of thought *typ* constants have properties (this is how we encode defaults) and are subject to manipulation in the reasoning process. That they are not "real" objects forces us to treat them somewhat judiciously and not altogether like actual objects, but they are objects to reason about nonetheless.

Enriching a formal reasoner's ontology to accommodate mental notions or concepts in general is not a new idea, nor is it new to AI ([Rapaport, 1981],[Rapaport, 1986] and [McCarthy, 1979] are examples). Indeed Brachman[Brachman, 1985] discusses the very idea of reifying typical mental notions to represent defaults in frame-like inheritance networks, yet no one seems to have carried this out in a logic-based approach to see what, if any, advantages and/or shortcomings might arise. One of Brachman's objections to the treatment of typicality in frame-like approaches is that the interpretation of the typicality concepts themselves is open to confusion and debate. There is no such confusion in the formalism we present here; a *typ* constant is, indeed, a "prototypical *individual* that somehow typifies the kind" ([Brachman, 1985] p 89.)

We should point out that we are not taking a detailed stand on the psychological nature or properties of mental concepts, except to grant that they are highly idiosyncratic. Rather, the very fact that people can conjure up, discuss, and reason about what, for them, is a typical tree, dog, or book, etc., lends some degree of cognitive plausibility to the approach introduced here, and to a large degree cognitive plausibility has motivated the efforts we report on.[1] Plausibility, of course, is fine as a motivator but the real test is whether or not *typ* constants exhibit desirable logical and computational properties.

As it turns out the *typ* constant formalism enjoys some very nice computational features which we discuss in section 7.2.1 after presenting the formalism itself in section 7.2. Then in section 7.3 we will discuss a severe limitation of the formalism which serves to illuminate a representational issue that has significance beyond our framework; namely the inability to express what we call *range* defaults. To our knowledge this more general form of the traditional default has thus far gone undetected in the literature. We conclude the paper with a discussion of the importance of *range* defaults and a brief exploration into their representation.

[1] Furthermore, truly robust commonsense reasoning agents ought to have the ability to reason about inexistants, e.g., Santa Claus, unicorns, and the "golden mountain", in general; mental typicality concepts are a special subclass of inexistants.

7.2 The Theory

We begin with a standard FOL with language $\mathcal{L}$. By way of notational convention, let Φ, Ψ, and various subscripted Φ_i's stand for formulae of $\mathcal{L}$ containing one free variable. We extend $\mathcal{L}$ to include one new constant symbol, typ_Φ, for each Φ in $\mathcal{L}$. We will call this extended language $\mathcal{L}'$ and the symbol 'a' will be used to stand for a closed term in $\mathcal{L}'$.

As alluded to above, the intended interpretation of the constant typ_Φ is the reasoner's notion of the typical Φ-entity so, for example, if Φ is $Bird(x)$ then typ_Φ is $typ_{Bird(x)}$; the (reasoner's) typical bird notion. (Note: From here on, for convenience, we drop the '(x)' from the typ constant notation and write, say, typ_{Bird} instead of $typ_{Bird(x)}$). Defaults, then, are encoded in expressions of the form $\Phi(typ_\Psi)$ which can be read "the typical Ψ is a Φ"(e.g., the default that "the typical bird flies" (or "typically birds fly") is encoded by $Fly(typ_{Bird})$). Sometimes we will write '$\Psi \stackrel{typ}{\rightarrow} \Phi$' instead of '$\Phi(typ_\Psi)$' (notice the transpositioning of 'Φ' and 'Ψ') because it is visually closer to the traditional way of depicting defaults and, hence, easier to read.[2] The '$\stackrel{typ}{\rightarrow}$' style is merely a notational convenience; '$\stackrel{typ}{\rightarrow}$' is not a new logical connective.

One proper axiom schema, **A**, and one (default) inference rule schema, **D** are added to the logic as follows: For all formulae Φ and Ψ (containing one free variable) and closed terms a of $\mathcal{L}'$

$$\mathbf{A}\text{: } \Phi(typ_\Phi)$$

$$\mathbf{D}\text{: } \frac{\Phi(typ_\Psi),\ \Psi(a),\ \text{Unknown } \neg\Phi(a)}{\Phi(a)}$$

Axiom schema **A** just assures that Φ applies to (typ_Φ); e.g., that the typical bird is a bird (i.e., $Bird(typ_{Bird})$), the typical singing bird sings and is a bird, i.e.,

$$Sings(typ_{Sings \wedge Bird}) \ \wedge \ Bird(typ_{Sings \wedge Bird})$$

and so on.

Rule schema **D** sanctions the judicious use of the encoded defaults which appear as the left most component to the antecedent of the rule. The intuition behind the rule is this: if a

[2] On the other hand, uncovering the *range* defaults that we discuss in section 7.3 was due, in part, to the original notation.

is a Ψ-thing then assume it to be as much like the typical Ψ-entity as possible. Thus if the typical Ψ-entity has property Φ then so too should a, unless known to the contrary. Bear in mind that 'a' in the rule is not confined to the original language, $\mathcal{L}$, but rather can be any term in $\mathcal{L}'$, and in particular may itself be a *typ* constant. This fact will be exploited in the next section when we show how to combine the encoded defaults to create encodings of new (i.e., newly formed) defaults.

The condition "Unknown $\neg\Phi(a)$" attached to schema **D** represents a criterion that tests for the appropriate application of the encoded default, thereby giving the formalism its nonmonotonic flavor. For the purposes of this paper we need not choose a particular implementation of "*Unknown*" though several possibilities come to mind, including fix-point consistency checking in the style of Reiter's Default Logic (DL) and McDermott and Doyle's NML (both undecidable), circumscription (semi-decidable) [3], and the negative introspection facility of step-logics (decidable) [Elgot-Drapkin, 1988] and [Elgot-Drapkin and Perlis, 1990].

7.2.1 Features

One nice feature of the *typ* constant approach to default reasoning is that, like circumscription, it requires no special logical connective (e.g., the '$\Rightarrow$' connective of [Delgrande, 1987], [Delgrande, 1988]) which is semantically distinguished from first-order material implication in order to write defaults. (Recall that '$\stackrel{typ}{\rightarrow}$' is only a notational convenience.) Nor are we committed to a modal operator that loosely corresponds to our "*Unknown*" (e.g., the 'M' operator of McDermott and Doyle[McDermott and Doyle, 1980]).

Delgrande helps point out another nice feature of our formalism by distinguishing the ability to reason *with* defaults from the ability to reason *about* defaults. Reasoning with a default is simply using it to come to (or not) a conclusion about an individual instance of the default rule. Thus we conclude that "Tweety" flies because she is a bird and birds typically fly. Default formalisms are geared toward this kind of reasoning and hence have a fair amount of success with it. Reasoning about defaults, on the other hand, essentially amounts to a reasoner's automatic ability to update its default database given some starting set of defaults. This, as Delgrande observes, is in general beyond the reach of DL and NML. For example, DL is unable

[3] In NML and circumscription, **D** would be written as an axiom schema rather than an inference rule.

to reason from "typically birds fly" and "typically flying things have wings" to "typically birds have wings". To do so would require a mechanism within the logic that automatically adds new rules of inference to a default theory since defaults in DL are themselves inference rules and not part of the logical language; DL has no such mechanism.

The theory presented here is able to model both reasoning with and about defaults (as can Delgrande's formalism and circumscription) in a highly desirable and natural way. To see this recall that the symbol 'a' in rule **D** may stand for either (1) a general term that is not a *typ* constant or (2) a constant symbol of the form typ_{Φ_1}. In the first case rule **D** operates, in spirit, much like a default rule in DL whereby a "real" domain object can be attributed a property by default. So, for example, $Flies(tweety)$ follows from $Flies(typ_{Bird})$, $Bird(tweety)$, and $Unknown(\neg Flies(tweety))$. In case (2) where '$a$' is typ_{Φ_1}, defaults can be combined to create new defaults. For example $Winged(typ_{Bird})$ follows from $Flies(typ_{Bird})$, $Winged(typ_{Flies})$, and $Unknown(\neg Winged(typ_{Bird}))$.[4] More generally, the following results characterize the way that the our theory combines defaults:

Theorem 7.1 (Transitivity Theorem for '*typ*')
If $\vdash \Phi_2(typ_{\Phi_1})$, $\vdash \Phi_3(typ_{\Phi_2})$, and $\vdash Unknown(\neg\Phi_3(typ_{\Phi_1}))$, then $\vdash \Phi_3(typ_{\Phi_1})$.
Note: Transitivity is easier to see when the theorem is written as:
$\vdash \Phi_1 \stackrel{typ}{\rightarrow} \Phi_2, \vdash \Phi_2 \stackrel{typ}{\rightarrow} \Phi_3, \vdash Unknown\neg(\Phi_1 \stackrel{typ}{\rightarrow} \Phi_3) \Rightarrow \vdash \Phi_1 \stackrel{typ}{\rightarrow} \Phi_3$

Proof: Using the default rule **D** (above) where Φ is Φ_3, Ψ is Φ_2, and a is typ_{Φ_1} the result follows immediately. ■

Theorem 7.2 (Compositionality Theorem for '*typ*' and '$\rightarrow$')

(i) If $\vdash \Phi_2(typ_{\Phi_1})$, $\vdash \forall x[\Phi_2(x) \rightarrow \Phi_3(x)]$, then $\vdash \Phi_3(typ_{\Phi_1})$.
(Alternatively: $\vdash \Phi_1 \stackrel{typ}{\rightarrow} \Phi_2, \vdash \Phi_2 \rightarrow \Phi_3 \Rightarrow \vdash \Phi_1 \stackrel{typ}{\rightarrow} \Phi_3$)

(ii) If $\vdash \forall x[\Phi_1(x) \rightarrow \Phi_2(x)]$, $\vdash \Phi_3(typ_{\Phi_2})$, and $\vdash Unknown(\neg\Phi_3(typ_{\Phi_1}))$ then $\vdash \Phi_3(typ_{\Phi_1})$.
(Alternatively, $\vdash \Phi_1 \rightarrow \Phi_2, \vdash \Phi_2 \stackrel{typ}{\rightarrow} \Phi_3, Unknown\neg(\Phi_1 \stackrel{typ}{\rightarrow} \Phi_3) \Rightarrow \vdash \Phi_1 \stackrel{typ}{\rightarrow} \Phi_3$)

Proof:

[4] In our more suggestive notation, $(Bird \stackrel{typ}{\rightarrow} Winged)$ follows from $(Bird \stackrel{typ}{\rightarrow} Flies)$, $(Flies \stackrel{typ}{\rightarrow} Winged)$, and $Unknown\neg(Bird \stackrel{typ}{\rightarrow} Winged)$.

(i) Follows immediately using substitution and modus ponens.

(ii) $\vdash \Phi_1(typ_{\Phi_1})$ from axiom **A**. $\vdash \Phi_2(typ_{\Phi_1})$ then follows from part (i) of the compositionality theorem. By the transitivity theorem, $\vdash \Phi_3(typ_{\Phi_1})$. ■

7.2.2 Some Observations

Much like circumscription, the bare theory presented here cannot adjudicate between prioritized competing or interacting defaults. Some additional machinery would be required to give intuitive results when an ordering relation (be it subset-superset, chronological, etc.) exists between interacting defaults, but we see no principled reason to suspect that this cannot be done.

On the other hand, one feature (among others) that distinguishes the present theory from circumscription (when the predicate *ab* is used to write defaults) is the subtle difference in the way the two formalisms treat "transitivity". In our approach, defaults combine to reflect the truly transitive reasoning: from "typically P's are Q's" and "typically Q's are R's" conclude "typically P's are R's". On the other hand, circumscription's *ab*-default representation produces something like "typical P's which are typical Q's are R's" from the same initial defaults. More precisely, consider the circumscription *ab*-defaults (1)–(4) below:

(1) $\forall x[(P(x) \wedge \neg ab(x, aspect_1)) \rightarrow Q(x)]$

(2) $\forall x[(Q(x) \wedge \neg ab(x, aspect_2)) \rightarrow R(x)]$

(3) $\forall x[(P(x) \wedge \neg ab(x, aspect_1) \wedge \neg ab(x, aspect_2)) \rightarrow R(x)]$

(4) $\forall x[(P(x) \wedge \neg ab(x, aspect_3)) \rightarrow R(x)]$

Default (3) follows from (1) and (2), but (4) does not (where '$aspect_3$' is a new aspect relating P and R). The difference between the two is that (3) reflects a more "cautious" sort of default composition than (4). Cautious because it "notes" the defaults that are being composed by conjoining the *aspects* that appear in them. Default (4) is what a truly transitive circumscriptive reasoner would produce given (1) and (2) and is, indeed, an analog of the result given in Theorem 1. We mention the distinction to point out that *typ* constants, in and of themselves, are not the cause of the discrepancy; if one prefers, a cautious reasoner can be modeled using *typ* constants by replacing rule **D** by:

$$\mathbf{D}'\text{: } \frac{\Phi(typ_\Psi),\ \Psi(a),\ \text{Unknown } \neg\Phi(a')}{\Phi(a')}$$

where $a' = typ_{\Phi\wedge\Phi_1}$ if a is of the form typ_{Φ_1}; $a' = a$ otherwise.

7.3 An Apparent Weakness Spawns a New Idea: Range Defaults

Very often, perhaps even most often, the mental concepts which correspond to *typ* constants either have or do not have (i.e., have the negation of) a given property. My typical person notion has two eyes, two arms, and a mouth; it is not an infant (i.e., it is a non-infant), nor is it the President of the United States. For me then, typ_{Person} is partially characterized by $HasTwoEyes(typ_{Person}), \ldots, \neg PresidentOfTheUS(typ_{Person})$. But some attributes of typical mental notions (and hence of *typ* constants) are not as well behaved. Again to use my typical person notion as an example, I think of it as being singularly gendered; it is male or female, not neuter and not hermaphroditic, yet it has no specified gender. There are too many male and too many female people to exclude either male-ness or female-ness as a possibility from my typical person notion. Moreover I know that both of the defaults "typically people are male" and "typically people are female" are too restrictive and hence are inappropriate. In short I have a *range* of possible default conclusions that can be drawn regarding the typical person and that range is maximally determined; it cannot be restricted any further.[5]

Perhaps the most general form of what we call *range* or *irreducible disjunctive* defaults is partly obscured in the example above by the fact that gender is normally thought of as being exhausted by $Male$ and $Female$ (i.e., $Male \leftrightarrow \neg Female$). In fact any number of properties, exhaustive or not, can be the range of a default. Consider this example: If you are anything like me your typical wood-notion has a color, it is either light brown (like pine) or dark brown (like walnut), but you can't pin it down any more than that. Clearly these color choices do not exhaust the possibilities for wood since there is black, pink, and even purple(!) wood. Not only do you believe that most wood is light brown or dark brown, you also believe that you cannot narrow or restrict this range of your default about wood color. There is just too

[5] Compare to [Levesque, 1986], where the opposite phenomenon is discussed.

much light brown wood to exclude light brown from the range of the default; similarly for dark brown.

How can these range defaults be faithfully represented in our theory? The short answer is that they can't because the theory has no way to distinguish between the negation of a default (which is itself a default) and the assertion that that default does not obtain. More precisely if we let $\Phi_1, \ldots, \Phi_n$[6] be the range (in the sense of the above example about wood) of a default about Ψ then we denote a range default by:[7]

$$\Psi \stackrel{typ}{\rightarrow} \boxed{\Phi_1 \vee \ldots \vee \Phi_n} \tag{7.1}$$

By 7.1 we intend, firstly, that:

$$\Psi \stackrel{typ}{\rightarrow} \Phi_1 \vee \ldots \vee \Phi_n \tag{7.2}$$

holds, and secondly that the range cannot be restricted any further. That is, for every *proper* subset, $\{\Phi_i, \ldots, \Phi_j\}$, of $\{\Phi_1, \ldots, \Phi_n\}$ the set of sentences given by:

$$\Psi \stackrel{typ}{\rightarrow} \Phi_i \vee \ldots \vee \Phi_j \tag{7.3}$$

are rejected as defaults. Let us introduce a new predicate letter, **INAP** (for "inappropriate"), and assume a quotation mechanism to reify wffs (in the style of [Perlis, 1985])[8] to try to express this latter condition. What we desire is to assert:

$$INAP(\Psi \stackrel{typ}{\rightarrow} \Phi_i \vee \ldots \vee \Phi_j) \tag{7.4}$$

(for every proper subset, $\{\Phi_i, \ldots, \Phi_j\}$, of $\{\Phi_1, \ldots, \Phi_n\}$) which is intended to mean that the defaults of 7.3 do not obtain. This we might try to express by an axiom such as:

$$INAP(\Psi \stackrel{typ}{\rightarrow} \Phi_i \vee \ldots \vee \Phi_j) \rightarrow \neg(\Phi_i \vee \ldots \vee \Phi_j)(typ_\Psi) \tag{7.5}$$

But now we get a contradiction which is most easily seen in the case where the n of sentence 7.2 is equal to 2 and the axiom $\forall x[\Phi_1(x) \leftrightarrow \neg\Phi_2(x)]$ holds. In that case sentence 7.4 amounts to:

$$INAP(\Phi_1) \wedge INAP(\Phi_2) \tag{7.6}$$

[6] We intend that no Φ_i is a subset of a disjunct of any others.

[7] The box notation is not part of the formal language.

[8] We will not show the quotation marks explicitly.

which together with 7.5 gives:

$$\neg\Phi_1 \wedge \neg\Phi_2 \quad \text{(i.e., } \Phi_1 \wedge \neg\Phi_1\text{)} \tag{7.7}$$

7.3.1 Why Bother?

Why bother with range defaults or, more precisely, with $INAP$ at all? Why not just represent the default 7.1 by 7.2 and ignore marking 7.3 as inappropriate? After all, 7.2 will give the right result for any instance of a Ψ-thing. That is, knowing that a particular chair is made of wood and that typically wood is light brown or dark brown would lead one to conclude (unless she knows to the contrary) that the chair is light brown or dark brown; no more, no less. A reasoner need not block the defaults "typically wood is light brown" and "typically wood is dark brown" to reach this conclusion. Likewise, in our formal theory the appropriate disjunction (e.g., $LightBrown(chair) \vee DarkBrown(chair)$) can be proven given the proper disjunctive default, and the overrestrictive conclusion(s) (e.g., $LightBrown(chair)$) will not be proven.

The answer is that there are cases in commonsense reasoning where it is not only important to reach the correct default conclusion, but also to have meta-knowledge about one's own defaults which itself can be reasoned with and about. As an example consider this: you know that cardinals are typically either red or russet in color but you can't pin it down any further since there are so many of each of the two colors.[9] Suppose you look out into the back yard of your house and notice that many cardinals, but only red ones, have gathered to eat. The simple default that typically cardinals are red or russet does not lead to the conclusion that the collection of birds in your back yard is in any way unusual. But you may have excellent reason to think it is an oddity because you have additional information, namely that it is *not* the case that typically cardinals are red. You can use this observation, that an unusual collection of birds has gathered in your yard, to wonder: Why have only red cardinals gathered? Do the russet cardinals not like the trees in the yard? And so on.

The above sort of knowledge that one may have about cardinals is precisely what a range default about cardinal color expresses, and this knowledge is crucial to the reasoning illustrated. Thus, not only is the formal representation of range defaults of interest in a purely

[9] Males are red, females russet.

theoretical sense (Can a formalism represent them?), it also has pragmatic ramifications for robust commonsense reasoning formalisms.

7.3.2 Other Formalisms and Range Defaults

The difficulty in representing range defaults is not peculiar to the *typ* constant approach. There is no way to write meta-level assertions about defaults in DL since, as we noted earlier, DL's defaults are inference rules and not part of the language, and thus there is no way to express that defaults are inappropriate. Circumscription, the most widely studied default mechanism, appears to have its own problems with range defaults. At first blush it seems that *ab*-defaults can be used in a rather straightforward way to get the desired representation. For example the range default about wood color might be expressed by the following three expressions:

$$\forall x[(W(x) \wedge \neg ab(x, aspect_1)) \rightarrow (LB(x) \vee DB(x))] \tag{7.8}$$

$$\neg\forall x[(W(x) \wedge \neg ab(x, aspect_2)) \rightarrow LB(x)] \tag{7.9}$$

$$\neg\forall x[(W(x) \wedge \neg ab(x, aspect_3)) \rightarrow DB(x)] \tag{7.10}$$

where LB, DB, and W stand for "light brown","dark brown", and "wood", respectively. But this is not without cost[10] . Sentence 7.8 is fairly straightforward and needs no explanation; 7.9 and 7.10, however, are not so easily dismissed. The mere appearance of $aspect_2$ and $aspect_3$ is somewhat unintuitive. What is their role? In a default rule an aspect serves as a bridge between two predicates to keep the effects of abnormality in check. But 7.9 and 7.10 are not defaults, rather they are the negations of defaults and their associated *aspects* seem unwarranted.

Regardless of the need for these *aspects* another difficulty arises when 7.8–7.10 are taken together together with, say, the wff "$W(chair)$". We would expect to get:

$$LB(chair) \vee DB(chair) \tag{7.11}$$

but 7.9 and 7.10 are counterexample axioms to 7.8 and [Perlis, 1986] shows that 7.11 will not be proven without some additional machinery (such as the scoping mechanism of [Etherington *et al.*, 1990]).

[10] Ignoring the fact that $2^n - 1$ of these *aspects* would need to be introduced for each range default, where n is the number of disjuncts in the range default.

7.4 Discussion

The representation of the denial of default information is an important part of commonsense reasoning. Such representation is not a simple matter of negating traditional representations, nor of *typ*-constant representations.

Is there a way to express range defaults using *typ* constants? Yes. But it appears we must modify our theory to do so. One way is to add a new inference rule relating $INAP$ to the encoded defaults and, additionally, alter rule **D** to accommodate expressions of the form $INAP(\alpha)$.[11] Another way, that we are currently exploring, is to reinterpret *typ* constants as sets of properties, *typ* sets, that apply to a typical mental notion. Thus typ_Ψ may or may not contain Φ and/or its negation, $\neg\Phi$. In particular, $INAP(\Phi(typ_\Psi))$ *iff* $\Phi \notin typ_\Psi$.

[11] The inference rule we have in mind is: from $\Phi(typ_\Psi)$ infer $\neg INAP(\Phi(typ_\Psi))$. The modification to **D** requires the addition of Unknown $INAP(\Phi(a))$ to the antecedent of the rule.

Epilogue (Summary and Future Work)

As stated at the outset of this dissertation our concern here has been with the general topic of reasoned change in belief. We have looked at belief change specifically as it relates to: (i) terminological change where new terms and new meanings may become important to a reasoner over time, particularly in the context of mistaken past beliefs; (ii) the changing meaning of the pronominal indexical "I"; and (iii) the problem of representing the denial of default information with an eye toward an agent's need to change her defaults (i.e., deny one default and accept another in its place) as she comes to learn more about the world.

This has been a representational effort; the particular kinds of tasks for belief change studied here seem to require for their formalization a rather formidable collection of predicates. We have presented formal treatments of: (i) **TU** in the context of the indexical "I" in FOL. (ii) **Distr** and **Mstkn** in a new active step-logic which can, under certain specified conditions, recover from contradictions. (iii) **FITB**, **RTA**, and **MISID**, in in the context of of objects of presentation, specifically with regard to object-identification errors, again in step-logic.

At the start we also indicated that this work can be viewed as a series of inroads toward accomplishing the long range goal of building a sophisticated reasoning system which has abilities suited to an advice-taker. Such a system, it is hoped, will accept advice from outsiders about its own beliefs – be it about the inappropriateness of a default or about an object-identification error – and then reason through to more appropriate beliefs. Let us take a brief look at some of the issues that arise out of this work which will need to be addressed in order to continue along the path toward this goal.

- The most glaring single gap in this treatment of reasoned belief change is the lack of a semantics in the terminological change and other related step-logic work. This is difficult because of the presence of contradictions. However, recent work on semantics of

contradictions (though not in a step-logic setting) may be relevant [Grant and Subrahmanian, 1992]. Another unusual feature for any potential semantics is the self-referential character of the logic, e.g., references to the logic's own history.

- Seeking cognitive verification of both the *short-chain* and *lazy-corroboration* hypotheses (chapter 4) are empirical endeavors. If either, or both, should turn out to be incorrect then our method of tracing through belief derivations (all-at-once *dc*-recovery) should be altered to account for the computational inefficiency which may arise in searching through prohibitively bulky derivations. A one-step-at-a-time approach might keep track, only of a theorem's most recent justifying beliefs, not the entire chain of reasoning. For instance a new rule for modus ponens might look like:

$$\frac{\mathbf{i}: \quad \alpha[S_1],\ \alpha \rightarrow \beta[S_2]}{\mathbf{i+1}: \quad \beta[\alpha, \alpha \rightarrow \beta]}$$

Here β's derivation contains α and $\alpha \rightarrow \beta$ to the exclusion of S_1 and S_2. A mechanism must be provided which, in response to the agent's distrust of a former belief, will trace through these limited derivations, step-by-step as reasoning progresses. It seems that step-logic's inherent step-wise nature will be amenable to such an approach, though a mindful eye should beware of circular derivations, e.g., as in:

$$Obs_{incons}(j) = \begin{cases} P, P \rightarrow Q_1, \ldots, Q_{n-1} \rightarrow Q_n, Q_n \rightarrow P & \text{if } j = 1 \\ \neg P & \text{if } j = k \\ \emptyset & \text{otherwise} \end{cases}$$

for a fixed k. The details and a related recovery theorem are left for future research.

- One concern which arises in regard to implementing a step-logic directly from its logical description is that of space. In robust domains the number of observations can be arbitrarily large, and so too can the number of theorems appearing at any step. This in turn creates a computational speed problem. Applying each inference rule in every way possible to all of a step's theorems may take much too long to be practical in most real-time problem solving or planning domains.

These concerns can be attacked on at least two fronts: One is to make the applicability of inference rules context-, focus-of-attention-, or goal-dependent, so that rules are only considered for application under the appropriate circumstances.

The other is to somehow inhibit the inheritance of beliefs, the primary offender of space consumption, allowing for reinstatement (in a more general form than discussed in chapters 4 and 5) when necessary. Reinstatement may then rely on some sort of relevance mechanism.[12] One way to restrict inheritance is to associate a *decay value*, d, with each theorem. The decay value of theorems which persist from one step to the next solely in virtue of inheritance will diminish as if the belief represented by the theorem is eventually pushed aside, or into "long-term memory", to make room for others. The time (i.e., step) may come when a theorem no longer appears, and once this happens it is not available for use in inference. But it may be possible to retrieve these older beliefs (i.e., make them current theorems once again) if and when the agent becomes newly interested in them. Theorems proven by other means, e.g., modus ponens, will have their decay values enhanced as they are likely to be currently relevant to the reasoner.

A preliminary start in this direction are the following inference rules. Here *max* stands for some maximum decay value (chosen by the designer of the logic); d, d_1, and d_2 are decay values associated with the given wffs at the specified step.

OBS_{dk}

$$\frac{\mathbf{i}: \quad}{\mathbf{i+1}: \quad \alpha \ \|max\|}$$

INH_{dk}

$$\frac{\mathbf{i}: \quad \alpha \ \|d\|}{\mathbf{i+1}: \quad \alpha \ \|d-1\|} \quad \text{where } d > 0$$

MP_{dk}

$$\frac{\mathbf{i}: \quad \alpha \ \|d_1\|, \ \alpha \rightarrow \beta \ \|d_2\|}{\mathbf{i+1}: \quad \beta \ \|max\|} \quad \text{where } d_1, d_2 > 0$$

[12] A idea akin to this for fixed-sized set of i-theorems (called STM) has been explored in the memory model of [Elgot-Drapkin *et al.*, 1987].

Once beliefs "decay out of memory" they can be stored in some efficient manner, say in a hash-table, and retrieved when relevant. If we represent the storage data-structure by the predicate **Stored**, then the rule RET_{dk} simulates the retrieval of $\alpha \rightarrow \beta$, a presumably relevant belief when α is an i-theorem.

$$RET_{dk} \qquad \begin{array}{ll} \mathbf{i}: & \alpha \, \|d_1\|, \; Stored(\alpha \rightarrow \beta) \, \|d_2\| \\ \hline \mathbf{i+1}: & \alpha \rightarrow \beta \|max\| \end{array} \qquad \text{where } d_1, d_2 > 0$$

These approaches only serve to retard the growth of the number of represented beliefs as time and reasoning progress, not to stall it completely. But if the limited-corroboration hypothesis proves to be correct, a plausible account of relevance, focus of attention, and decay are worked out, and an an efficient treatment of *Stored* is offered – all major endeavors – then these suggestions might offer some promise.

- Our treatment of *presentations* (chapter 3) has been extremely informal. But a deeper understanding of term-change (and of the factors present in the other cognitive tasks as well) may result from a careful study of presentations and perception-based beliefs.

- We assume the correct *reality terms* are introduced through by tutors, though this is not always realistic. We would like to formalize a hypothesize-and-test process based on our *reality terms* that enables an agent to speculate about, and perhaps uncover, his own mistakes.

- The *Two Johns* problem can be generalized to the *N Johns* problem in which an agent's belief set represents the conflation of N different people named "John". The *Mistaken Car* problem can be generalized to the *Mistaken Objects* problem wherein an agent's belief (or set of beliefs) reflects multiple object-identification errors at one time. Both generalizations are in need of exploration.

- Gilbert's Spinozist analysis of comprehension, acceptance, and rejection of beliefs [Gilbert, 1991] may turn out to have deeper ties with various themes here. In particular, the notion

of past beliefs that are *not* viewed as mistaken but that are viewed as open to question, may be related to Spinozist comprehension. Also, there may be another "mode" of belief, namely "tentative acceptance" that does not commit the reasoner to action based on such a belief but is more like "trying a belief on for size". Such a mode would resemble acceptance insofar as *reason* goes, but not action.

Bibliography

[Allen and Perrault, 1980] Allen, J. and Perrault, R. 1980. Analyzing intentions in utterances. *Artificial Intelligence*, 19:143–178.

[Allen, 1984] Allen, J. 1984. Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154.

[Astington and Gopnik, 1988] Astington, J. W. and Gopnik, A. 1988. Knowing you've changed your mind: Children's understanding of representational change. In Astington, J. W., Harris, P. L., and Olson, D. R., editors, *Developing Theories of Mind*, pages 193–206. Cambridge University Press.

[Astington *et al.*, 1988] Astington, J. W., Harris, P. L., and Olson, D. R., editors. 1988. *Developing Theories of Mind*. Cambridge University Press.

[Barwise and Perry, 1983] Barwise, J. and Perry, J. 1983. *Situations and Attitudes*. MIT Press, Cambridge, MA.

[Brachman, 1985] Brachman, R. 1985. I lied about the trees or, defaults and definitions in knowledge representaion. *AI Magazine*, 6(3):80–93.

[Cummins, 1989] Cummins, R. 1989. *Meaning and Mental Representation*. MIT Press, Cambridge, MA.

[da Costa, 1974] da Costa, N. 1974. On the theory of inconsistent formal systems. *Notre Dame Journal of Formal Logic*, 15:497–509.

[deKleer, 1986] deKleer, J. 1986. An assumption-based TMS. *Artificial Intelligence*, 28:127–162.

[Delgrande, 1987] Delgrande, J. P. 1987. A first-order conditional logic for prototypical properties. *Artificial Intelligence*, 33(1):105–130.

[Delgrande, 1988] Delgrande, J. P. 1988. An approach to default reasoning based on first-order conditional logic: Revised report. *Artificial Intelligence*, 36(1):63–90.

[Doyle, 1979] Doyle, J. 1979. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272.

[Elgot-Drapkin and Perlis, 1990] Elgot-Drapkin, J. and Perlis, D. 1990. Reasoning situated in time I: Basic concepts. *Journal of Experimental and Theoretical Artificial Intelligence*, 2(1):75–98.

[Elgot-Drapkin *et al.*, 1987] Elgot-Drapkin, J., Miller, M., and Perlis, D. 1987. Life on a desert island: Ongoing work on real-time reasoning. In Brown, F. M., editor, *Proceedings of the 1987 Workshop on The Frame Problem*, pages 349–357. Morgan Kaufmann. Lawrence, Kansas.

[Elgot-Drapkin, 1988] Elgot-Drapkin, J. 1988. *Step-logic: Reasoning Situated in Time*. PhD thesis, Department of Computer Science, University of Maryland, College Park, Maryland.

[Etherington *et al.*, 1990] Etherington, D., Kraus, S., and Perlis, D. 1990. Nonmonotonicity and the scope of reasoning: Preliminary report. In *Proceedings of the 8th National Conference on Artificial Intelligence*, pages 600–607, Boston, MA. AAAI.

[Etherington, 1988] Etherington, D. 1988. *Reasoning with Incomplete Information*. Research Notes in Artificial Intelligence. Morgan Kaufmann, Los Altos, CA.

[Flavell *et al.*, 1986] Flavell, J., Green, F., and Flavell, E. 1986. Development of knowledge about the appearance-reality distinction. *Society for Research in Child Development Monographs*, 51. No. 1, Series No. 212.

[Fodor, 1979] Fodor, J. 1979. *The Language of Thought*. Harvard University Press.

[Frege, 1956] Frege, G. 1956. The thought. *Mind*, 65:289–311.

[Gilbert, 1991] Gilbert, D. Feb. 1991. How mental systems believe. *American Psychologist*, 46(2):107–119.

[Grant and Subrahmanian, 1992] Grant, J. and Subrahmanian, V. 1992. The optimistic and cautious semantics for inconsistent knowledge bases. Presented at the Symposium on Logic in Databases, Knowledge, Representation, and Reasoning.

[Haas, 1986] Haas, A. 1986. A syntactic theory of belief and action. *Artificial Intelligence*, 28:245–292.

[Harman, 1986] Harman, G. 1986. *Change in View: Principles of Reasoning.* MIT Press.

[Harper and Charniak, 1986] Harper, M. and Charniak, E. 1986. Time and tense in English. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, Columbia University, New York.

[Hirst, 1981] Hirst, G. 1981. *Anaphora in Natural Language Understanding: A Survey.* Springer-Verlag, Berlin.

[Hirst, 1991] Hirst, G. 1991. Existence assumptions in knowledge representation. *Artificial Intelligence*, 49:199–242.

[Kripke, 1980] Kripke, S. 1980. *Naming and Necessity.* Harvard University Press.

[Leslie, 1987] Leslie, A. M. 1987. Pretense and representation: The origins of 'Theory of Mind'. *Psychological Review*, 94(4):412–426.

[Leslie, 1988] Leslie, A. M. 1988. Some implications of pretense for mechanisms underlying the child's theory of mind. In Astington, J. W., Harris, P. L., and Olson, D. R., editors, *Developing Theories of Mind*, pages 19–46. Cambridge University Press.

[Levesque, 1984] Levesque, H. 1984. A logic of implicit and explicit belief. In *Proceedings of the National Conference on Artificial Intelligence*, pages 198–202, Austin, TX. AAAI.

[Levesque, 1986] Levesque, H. 1986. Making believers out of computers. *Artificial Intelligence*, 30:81–108.

[Lin, 1987] Lin, F. 1987. Reasoning in the presence of inconsistency. In *Proceedings of the 6th National Conference on Artificial Intelligence*, Seattle, WA. AAAI.

[Maida, 1991] Maida, A. 1991. Maintaining mental models of agents who have existential misconceptions. *Artificial Intelligence*, 50(3):331–383.

[Maida, 1992] Maida, A. March 1992. Propositionally representing incomplete knowledge about existence. Stanford University. AAAI92 Spring Symposium Series on Propositional Knowledge Representation. To Appear.

[McCarthy and Lifschitz, 1987] McCarthy, J. and Lifschitz, V. 1987. Commentary on McDermott. *Computational Intelligence*, 3(3):196–197.

[McCarthy, 1958] McCarthy, J. 1958. Programs with common sense. In *Proceedings of the Symposium on the Mechanization of Thought Processes*, Teddington, England. National Physical Laboratory.

[McCarthy, 1979] McCarthy, J. 1979. First order theories of individual concepts and propositions. *Machine Intelligence*, 9:129–147.

[McCarthy, 1980] McCarthy, J. 1980. Circumscription: A form of non-monotonic reasoning. *Artificial Intelligence*, 13(1,2):27–39.

[McDermott and Doyle, 1980] McDermott, D. and Doyle, J. 1980. Non-monotonic logic I. *Artificial Intelligence*, 13(1,2):41–72.

[McDermott, 1982] McDermott, D. 1982. A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6.

[Miller and Perlis, 1987a] Miller, M. and Perlis, D. 1987. Proving facts about 'I'. In *Proceedings of the 10th Int'l Joint Conference on Artificial Intelligence*, pages 499–501.

[Miller and Perlis, 1987b] Miller, M. and Perlis, D. 1987. Proving self-utterances. *Journal of Automated Reasoning*, 3:329–338.

[Miller and Perlis, 1991] Miller, M. and Perlis, D. 1991. Typicality constants and range defaults: Some pros and cons of a cognitive model of default reasoning. In *Proceedings of the 1991 SIGART International Symposium on Methodologies for Intelligent Systems.*

[Miller and Perlis, 1993a] Miller, M. and Perlis, D. 1993. Presentations and *this* and *that*: Logic in action. In *Proceedings of the 15th Annual Meeting of the Cognitive Science Society*, Boulder, CO. To appear.

[Miller and Perlis, 1993b] Miller, M. and Perlis, D. 1993. What experts deny, novices must understand. To be presented at the *3rd International Workshop on Human and Machine Cognition*.

[Ohlbach, 1984] Ohlbach, H. 1984. Predicate logic hacker tricks. *Journal of Automated Reasoning*, 1:435–440.

[Perlis, 1985] Perlis, D. 1985. Languages with self reference I: Foundations. *Artificial Intelligence*, 25:301–322.

[Perlis, 1986] Perlis, D. 1986. On the consistency of commonsense reasoning. *Computational Intelligence*, 2:180–190.

[Perlis, 1988] Perlis, D. 1988. Languages with self reference II: Knowledge, belief, and modality. *Artificial Intelligence*, 34:179–212.

[Perlis, 1991] Perlis, D. 1991. Putting one's foot in one's head – part I: Why. *Noûs*, 25:325–332. Special issue on Artificial Intelligence and Cognitive Science.

[Perry, 1977] Perry, J. 1977. Frege on demonstratives. *The Philosophical Review*, 65(4):474–497.

[Perry, 1979] Perry, J. 1979. The problem of the essential indexical. *Nous*, 13:3–21.

[Priest and Routley, 1984] Priest, G. and Routley, R. 1984. Introduction: Paraconsistent logics. *Studia Logica*, 43:3–16.

[Rapaport, 1981] Rapaport, W. 1981. How to make the world fit our language: An essay in Meinongian semantics. *Grazer Philosophische Studien*, 14:1–21.

[Rapaport, 1986] Rapaport, W. 1986. Logical foundations for belief representation. *Cognitive Science*, 10:371–422.

[Reiter, 1978] Reiter, R. 1978. On Closed World Data Bases. In Gallaire, H. and Minker, J., editors, *Logic and Data Bases*, pages 55–76. Plenum Press, New York.

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

[Rieger, 1974] Rieger, C. 1974. *Conceptual Memory: A Theory and Computer Program for Processing the Meaning Content of Natural-Language Utterances.* PhD thesis, Department of Computer Science, Stanford University, Palo Alto, California.

[Smith, 1986] Smith, B. 1986. Varieties of self-reference. In *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning about Knowledge*, Los Altos, California. Morgan Kaufmann.

[Smullyan, 1978] Smullyan, R. 1978. *What is the Name of This Book?* Prentice-Hall, Englewood Cliffs, New Jersey.

[Wimmer and Perner, 1983] Wimmer, H. and Perner, J. 1983. Beliefs abut beliefs: Representation and constratining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13:103–128.