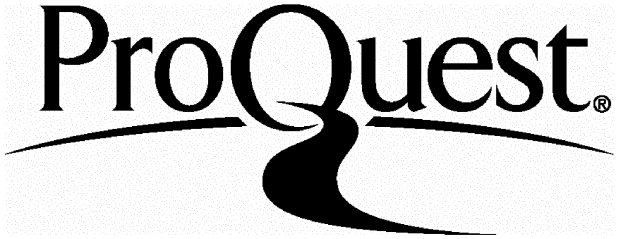
# Motivated Inductive Discovery

**Michael Mordechai Luck**

A dissertation submitted in partial fulfillment
of the requirements for the degree of
**Doctor of Philosophy**
of the
**University of London**

Department of Computer Science
University College London

1993



# Abstract

Research in machine discovery to date has tended to concentrate on the replication of particular episodes in the history of science, and more recently on the extraction of regularities from large databases. In this respect, current models of induction and discovery concentrate solely on the acquisition of knowledge, and lack the flexibility of reasoning that is necessary in a real-world changing environment.

Against this backdrop, this dissertation addresses inductive reasoning, specifically based around the scientific discovery paradigm. A framework for inductive reasoning is presented which includes the six stages of prediction, experimentation, observation, evaluation, revision and selection. Within this framework, different kinds of inductive reasoning can be reduced to the same basic component processes. The difference between the various kinds of reasoning arises not through the use of different mechanisms, but through the influence of motivations which bias the application of these mechanisms accordingly. Also within this framework, a model and its implementation as a computer program, the MID system, have been developed, concentrating primarily on the internal stages of the framework, prediction, evaluation, revision and selection. The role of motivations in allowing reasoning for both knowledge and action is investigated and implemented in the program. By choosing different internal models of motivation for reasoning systems, different behaviours can be achieved from the same basic architecture.

The MID system reasons in simple physical domains, both for knowledge and for action. It demonstrates how a basic mechanism can be used to provide an effective means for reasoning in a variety of contexts, and also how a simple motivational representation can be used as an effective control strategy.

# Acknowledgements

I could not have completed this thesis without the support of my supervisor, Derek Long. I thank him for his encouragement and patience, and for the very many hours spent in discussion and debate.

For their time reading and discussing various parts of this thesis with me, thanks must also go to Prof. Arthur Miller of the Department of the History and Philosophy of Science at UCL, Dr. Peter Cheng of Nottingham University, Alexandra Coddington, Maria Fox, Mark d'Inverno and Gordon Joly.

In addition, many people have helped by providing a stimulating environment in which to work, both academically and socially. In no particular order, I thank Paul Samet, Felicity Dams, Sara Schwartz, Sophia Prevezanou, Mark Levene, Suran Goonatilake, Carl Evans, John Wolstencroft, John Washbrook, Charles Easteal, David Lee, Dave Parrott, Stuart Clayman, Mark Jones, Simon Courtenage, Owen Mostyn-Owen, the UCL AI Group, Daniel Gordon, Sateen Bailur, Sabi Kabeli and Cindy Freedman.

Finally, I thank my parents, Harry and Dalia, and my sister, Sharon, for their love and support over the years.

This research was carried out under a Science and Engineering Research Council studentship.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

> The unity of all science consists alone in its method, not in its material.
>
> — Karl Pearson, *The Grammar of Science*

## 1.1 Introduction

Throughout history, much of human endeavour has been directed at increasing the knowledge available about the world. An important aim of science is, arguably, to increase our understanding of the world in order that we may explain and predict events as part of an ongoing effort to mitigate the effects of our environment. Such is the importance of knowledge and scientific progress that the nature of science as an activity in itself has also been studied extensively. The investigation of scientific reasoning is being pursued along a number of fronts, inspired by episodes in the history of science, and by the rewards that will be provided by a better understanding. Many different accounts of the nature of scientific activity have been suggested, ranging from philosophical attempts to define it logically through to sociological and historical analyses. More recently, artificial intelligence (AI) has provided techniques that allow scientific reasoning to be investigated computationally. This thesis is concerned with the development of a computational approach to what we call scientific reasoning.

Scientific reasoning, and in particular induction and discovery, can be applied not only in scientific domains and contexts, but also to more common situations. This thesis claims that it is possible to provide a common framework within which different varieties of inductive reasoning are unified, and from which a strong model of inductive reasoning can be developed. By breaking down the reasoning procedure into its component parts,

each component can be investigated separately and its role in the different kinds of reasoning considered. In doing this, problems are identified, and a flexible and robust model of reasoning that allows for these variations can be developed.

This chapter begins by discussing the role of knowledge in artificial intelligence and its associated problems, outlining some of the deficiencies of current 'intelligent' systems. It continues with a clarification of what is meant here by scientific reasoning, discovery, induction and other terms which have become confused over time. Some background is then introduced to provide a general perspective on the relation between this and other work. Finally, the aims and motivations of the work are discussed, and an overview of the thesis is presented.

## 1.2 The Role of Knowledge in Intelligence

The significance of knowledge in intelligence is undeniable. It is widely held that knowledge is the primary force behind any system that can exhibit intelligent understanding and action at a high level of competence (eg. [68]). If it is not the primary force, it is certainly a necessary force. Without knowledge, or even just with little or poor knowledge, the capability for intelligence is seriously curtailed. Currently, a number of research efforts (such as the CYC project [69]) are directed at encoding a large and varied body of knowledge in the belief that this will enable the construction of general intelligent machines. Expert systems demonstrate very effectively the capabilities of knowledge-based technology at one end of the artificial intelligence spectrum. The knowledge that is encoded within an expert system is typically limited to a small domain of application, however, but provides a useful and effective means for 'understanding' that domain.

Such knowledge-intensive techniques face many shortcomings, however. No matter how simple the computational machinery involved, the problem of knowledge acquisition cannot be avoided, and should not be underestimated. The difficulties of expertise elicitation and knowledge transfer, for example, are well known (see Gaines [25]), and suffice it to say here that knowledge-based systems must not only be judged on performance, but also on the work required to acquire that knowledge in the first instance. In addition, acquisition of expert knowledge requires that all forms of potential interactions with that knowledge be anticipated beforehand to ensure a proper functioning of systems that use it. Furthermore, the explicit encoding of knowledge imposes restrictions upon the con-

tent of the knowledge base that may later prove critical in any one of a number of tasks undertaken, including the ability to learn effectively.

Our understanding of what constitutes knowledge itself is problematic, but whatever notion of knowledge we may adopt, knowledge is ultimately dependent upon the changes that occur in our environment over time. Knowledge, in many ways, is in flux. In other words, what might be correct or consistent at one time might not be so at another. The world is a dynamic fluid system, which demands that any repository of knowledge be easily and efficiently modified so that it remains consistent with a changing reality. In a concrete, real world context, we can relate this to the changes in our environment which influence our everyday actions. For example, the knowledge that Margaret Thatcher is Prime Minister might be encoded, only to discover some months later that this is wrong, and that John Major is Prime Minister instead. (In reality, we know that the situation of any single person being Prime Minister is only temporary, so we should allow for the modification of that knowledge.)

Furthermore, knowledge, in a global sense, is not complete. Continually, we discover more and more about the world in which we live; we discover things that were not known before. This applies just as equally to scientific research which we can think of as *communal knowledge*, as to *individual knowledge* about our own individual environment. For example, advances in medicine (communal, scientific knowledge) have lead to a greatly decreased infant mortality rate. At an individual level, one might 'discover' that a tube of toothpaste is empty. In short, there is always the potential to add to knowledge, and we must make allowance for the addition of such newly-discovered knowledge to our knowledge bases.

Recent work in AI has begun to address these issues as a result of different lines of research. First, the problems associated with using static, fixed and rigid *knowledge bases* have been recognised, and the need for more flexible techniques allowing dynamic modifications to be made to such knowledge bases identified. Second, research in the philosophy of science, particularly that which is concerned with scientific discovery and induction, has been given a new impetus with the advent of computational techniques and, consequently, computational models of discovery are now being developed. Third, a move towards consideration of real-world problems and scenarios, and their associated variability, has necessitated a reappraisal of knowledge (and belief) as fluid and subject to change.

As progress continues in other areas of research concerned with using explicitly encoded knowledge, these issues are becoming ever more important, demanding the development of systems which are capable of effective knowledge management as an inherent part. Such capabilities will enable:

- The automatic generation of knowledge bases, avoiding the problems of knowledge acquisition with human experts.
- The speedy construction of prototype domain theories.
- The modification of incorrect or inconsistent knowledge, including the imperfections permitted by rapid prototyping.
- The addition of newly-discovered knowledge to existing domain theories.
- The maintenance of correct knowledge in a rapidly changing environment.

This more closely mirrors the way things work in real world situations, and provides a sound basis for learning systems.

## 1.3 Scientific Reasoning

### 1.3.1 Introduction

The above discussion identifies a number of problems that remain largely unsolved in AI. Classical logically valid reasoning techniques, primarily deduction (but also other reasoning methods), while having a definite role to play in artificial intelligence, are unsuitable here precisely because of their rigour. Deductive inferences are *explicative* in that they reveal the relationships in existing knowledge, and allow for transitions between small basic components and large complex structures. By contrast, what might loosely be called *scientific reasoning*, aims to create new knowledge, to extend the knowledge that may already exist. It is *scientific* in that it follows the aim of science in increasing knowledge about the world. It is complementary to formal logic, but since it does not lie in the realm of formal logic, it is not guaranteed to be correct or even to draw an inference at all. Scientific reasoning broadly encompasses induction and discovery techniques. These are considered below.

### 1.3.2 What is induction?

There are many different concepts of what constitutes induction, and many different levels of detail to that understanding. This is, in part, due to the different emphases that have been placed on it by a variety of diverse groups and individuals. Philosophical concerns with the logical (or otherwise) validity of induction may be different to those of computer science interested in achieving certain results, and both of these will be different from psychological concerns with induction which stem from understanding how it is used in human reasoning processes. Even within the same field, judgements and concepts vary to a great degree. A notable example is that of Mill who regarded induction as a logical procedure analogous to deduction in contrast to the vast majority of the philosophical community of the time. The continuing presence of heated debate and disagreement over the nature and role of induction is indicative of its significance. The ambiguity surrounding it and the lack of a consensus over definition embody the expressiveness that is inherent. Yet in order to discuss induction meaningfully, we must tie it down to definite ideas and procedures. Here, then, we aim for an informal yet clear description of what we mean by induction.

First of all it is important to draw the distinction between scientific induction, which concerns us here as a means for addressing the above issues, and mathematical induction, which is an entirely different matter. Scientific induction is so called because of its original invocation as a suitable reasoning method for science or for discovering knowledge, and because of the now dismissed claim that it provided a logically valid complement to deduction.

The view that science proceeds by inductively inferring laws *directly* from observations without intermediate hypotheses was always problematic, and is now discredited. In its place has arisen the notion of a methodology or programme for science rather than a rigorous logical procedure. Traditionally, such methods have avoided the problem of the creation of hypotheses in the first instance, and instead concentrated on the testing, and refutation or revision of hypotheses as appropriate. The *hypothetico-deductive method* which addresses these later stages of induction through logical analysis has been subjected to much criticism. We take a pragmatic position on this, recognising the role of elements of the hypothetico-deductive method in some form in scientific reasoning and also in everyday reasoning, and noting the power that it brings to computational models of such

reasoning.

A dictionary definition takes induction to be the process of inferring a general law or principle from the observation of instances. This is close, but requires a little modification.

**Definition** *Induction* is the process of inferring an explicit general conclusion primarily from observation of instances.

This allows the notion of inference and the kind of conclusion to be interpreted in a number of ways, but requires that the premises of an inductive argument are observations. It extends the scope of induction through to all domains and contexts, not just scientific ones.

By induction, then, we mean scientific induction as denoting reasoning that is based on empirical evidence obtained through observation of the world. Induction in this sense may thus be harnessed through a methodology for reasoning such as discovery. It can be seen to provide constraints on the nature of the reasoning laid out in a more precise and well defined system.

### 1.3.3 What is discovery?

As mentioned above, the notion of induction of laws *directly* from observations is inadequate. In response to this, a shift away from the notion of induction as a logical procedure introduced the concept of a system of scientific discovery for 'doing science'. Such systems set out rules of procedure for a programme designed to uncover laws and principles governing the nature of the world. Many programmes of discovery have been, and continue to be devised. Traditionally, these have been inductive, only admitting observations as a basis for reasoning, or at least excluding those parts of the programme which may suggest other influences, asserting that they are outside science. More recently, work on scientific reasoning has acknowledged the role of other factors, including such techniques as analogical reasoning, in the scientific process. Discovery is a broad notion that admits many factors and influences.

Discovery is usually restricted to science. This is a restriction on the reasoning process to the *communal knowledge* mentioned earlier, but there is no reason why it should not also apply to *individual* or non-scientific knowledge. Discovery is difficult to define because of disagreement about what it is that constitutes discovery, and how broad its scope should be [130]. We can define discovery as follows:

**Definition** *Discovery* is the process of finding out new knowledge.

This definition applies as easily to individual knowledge as it does to communal knowledge. What is known to one person may yet be discovered by another.

This thesis concentrates on *inductive discovery* — that is to say it is concerned primarily with discovery that is constrained by a reliance on empirical observations. The word *discovery* denotes the nature of the problem or the task at hand, while the word *inductive* denotes the kind of reasoning used to address it; inductive reasoning as opposed to analogical reasoning or any other. Thus we can define inductive discovery:

**Definition** *Inductive discovery* is the process of finding out new knowledge from observation of instances.

In this thesis, the terms *induction, discovery,* and *inductive discovery* will all be used to denote the same thing, discovery of the inductive kind, unless explicitly stated otherwise. Indeed, these terms are usually used to refer to the same kind of reasoning process, but in different contexts.

## 1.4 Perspectives on Induction and Discovery

As with much of AI, scientific reasoning has its roots deep in the history and philosophy of science. An aim of science can be thought of as the acquisition of knowledge through experimentation and observation of the world. Attempts to achieve a better understanding of nature have thus spawned many methodologies and programmes for science. Psychology, too, is intimately bound up with AI in the investigation of intelligence, with areas devoted to investigating and understanding human thought and reasoning processes.

The concern of this thesis is not with philosophical or psychological models or theories. Emphasis is placed firmly on a computational approach. It would be reckless, however, to ignore the vast amount of research that has been directed at the problems considered here from these alternative but complementary points of view. Indeed, the advent of the computer has provided a new impetus both to philosophical and psychological research on discovery and induction (see, for example, [116, 117]) which links up strongly with AI. Thus we can look to psychology, philosophy and other fields for inspiration towards solving many of the problems that confront us in AI. We might differentiate between philosophy and psychology by saying that the concern of psychology is with understanding these processes in humans (and animals) while the concern of philosophy is with developing valid and effective processes for achieving greater knowledge of the

world. The distinction is not firm, however, and in areas of cognitive science, for example, philosophy and psychology merge in some of these points. It can be argued that much of AI research follows on from long strands of research in the philosophy of science and psychology, and as such it is important to provide some background.

### 1.4.1 A Philosophical-Historical Perspective

Although the discussion and investigation of knowledge, and what is now known as science and the philosophy of science can be traced as far back as Plato and Aristotle, the usual starting point for a discussion of the work in this area is the Seventeenth Century. This is primarily due to two factors. First, the philosophers and scientists of the time believed that their work was something entirely different from what went before, although as has been pointed out [82], there are strong links to Aristotle and Plato. Second, the sudden and rapid advance of science in the Seventeenth Century, with scientists such as Galileo and Newton producing remarkable and significant results, provided a new impetus to investigating the question of how knowledge, scientific or otherwise, was acquired.

#### Early Empiricism and Naive Inductivism

Empiricism is usually defined as, "the thesis that all knowledge of matter of fact as distinct from that of purely logical relations, is based on experience [21]." Francis Bacon, an important forerunner of the empiricist tradition, was perhaps the first significant contributor to the methodology of science though he made no real contribution to science itself. He gave examples of the use of his new methodology which was intended to search for the causes of observed effects. Briefly, it involved the formulation of hypotheses, the consequences of which were then tested against new data. This would lead to the elimination of hypotheses which were incorrect, and eventually to the true explanation of the effect. However, it depended for its success on a wide base of empirical information.

The fact that Bacon made no significant contribution to science itself is important because a short while later, Isaac Newton, whose contribution to science concerning mechanics and optics was phenomenal, denied the use of hypotheses in his reasoning. He argued that certainty was required, and that it was to be achieved by reasoning inductively from experiments and observations alone. A belief in the uniformity of nature allowed the use of experimental 'proofs' and the deducibility of general conclusions from these observations. Whether or not Newton actually used hypotheses in his own reasoning

is a matter for debate. The important point here is that he claimed that hypotheses were neither necessary nor desirable for inductive reasoning.

His claim of direct inference of general laws from specific observations which might appropriately be called *naive inductivism* because of the lack of any intermediate hypotheses, became part of the *problem of induction.* This came to the fore with Hume (who formulated it as such) much later. On considering the matter of causality [44], the question was raised of whether or not it is reasonable to believe in the uniformity of nature, or whether there are ever grounds for believing that exact conclusions can be attained by an inductive argument. Hume, however, denied the principle of the uniformity of nature, giving a psychological account of our belief in it. Inductive generalizations are never justified. Yet Hume provided a set of rules for scientific inquiry, a methodology, despite his misgivings over causation and induction, and in other works he recommended one of Newton's rules of reasoning which embodied the essence of naive induction. This inconsistency seems to reveal some pragmatism, and an identification of the need to avoid paralysis of action.

### Logical Positivism

The empiricism of Hume and more contemporary empiricists provided a foundation for the very influential school of logical positivism (or logical empiricism) which was established in the first half of this century. The *empirical* component maintained that all knowledge must be grounded on experience. This was fixed in the verifiability principle which stated that the meaning of a proposition consists in the method of its verification, which is whatever observations (as experiences) show. Questions of theology and metaphysics are thus neither true nor false, but become meaningless and inadmissible as a consequence of their unverifiability. The *logical* aspect of the programme was intended to systematize science through the manipulation of empirical propositions using symbolic logic in an attempt to provide a formal rendering of its structure. Any proposition that is not observable (ie. theoretical) must thus be indirectly determined via observational propositions and the use of logic to specify the relationship between the two.

In the discussion of induction, the Logical Positivists made two important contributions. First, they distinguished between the context of discovery in which hypotheses were developed, and the context of justification in which they were assessed. The discovery of hypotheses was a problem that was left to psychologists to explain, since it was

considered that it might well be nonlogical. Second, the emphasis on verification led to the development of the notion of confirmation. They maintained that collecting positive evidence confirming a hypothesis should increase the confidence in its truth.

### Against Verification

Logical Positivism, in attempting to unite the rigour of logic with the epistemology of empiricism, admitted serious flaws. These were most effectively exposed by Karl Popper (among others), who proposed an alternative methodology for science [87]. In particular, the difficulty that general empirical statements cannot be verified because of the problem of induction was a major concern, and Popper attempted to avoid this by replacing the traditional concept of confirmation with falsification. *Falsificationism* is based on the fact that logic permits the establishment of the falsity but not the truth of theories in the light of observations. Science thus begins with problems for which falsifiable hypotheses are formulated as solutions. These hypotheses are then subjected to experimentation and criticism in the course of which some will be deductively refuted while others may remain. In the course of testing these hypotheses, the data collected may lead to new problems which will need to be accommodated. This leads to the introduction of new hypotheses which must, in turn, be tested. Popper contends that the continual application of this method of conjectures and refutations is the basis for the progress of science. A hypothesis is never regarded as being true even if it has passed a wide variety of stringent tests, but it may be considered superior to its predecessors.

There are a number of important points here. Like the Logical Positivists, Popper recognises two distinct phases in science, the imaginative phase as discovery and the critical phase as justification. He only considers the critical phase in his programme since he regards the invention of hypotheses as being irrational and outside science. Falsifiability is also used as a criterion for demarcation between science and non-science, those systems which are unfalsifiable such as astrology being deemed pseudo-science and unsuitable for reasoning, since they can never be refuted.

## 1.4.2 A Psychological Perspective

Psychological approaches to the problem of scientific discovery have been distinguished from others as involving analyses of the actual behaviour of humans engaged in aspects of scientific reasoning [126]. Klahr et al. [49] further divide the psychological approaches

into those which use a retrospective analysis of the scientific record of real scientists making real discoveries, and those which recreate simulated laboratory contexts for scientific discovery. The first of these can be considered an historical approach. The second has enabled a detailed analysis of the behaviour of subjects under highly controlled conditions, and an immediate investigation of the thought and reasoning processes involved. It does, however, suffer from the drawback that it is only analogous to science rather than being actual science. Nevertheless, there is a history of solid psychological research into induction and discovery, with concerns ranging from human acquisition of sequential patterns thirty years ago (eg. [132], [109],[50]) through to current efforts explicitly concerned with the nature of human scientific reasoning in more realistic discovery problems (eg. [49], [20], [99]). A recent review of much psychological research on discovery can be found in [33].

This work shares a concern with the manner in which people actually reason, but the emphasis here is not on modelling human cognition, but on developing effective techniques for scientific reasoning that exploit the capabilities of computers.

## 1.5 Aims and Motivation

Research is currently being carried out on many aspects of discovery in many forms from a variety of perspectives. Work is being done on theory revision, theory formation, theory choice, numerical discovery, and so on. All of these are relevant, yet the plethora of terms and apparently different paradigms has led to a fragmentation resulting in a collection of distinct parts. Important motivations of this research are the belief that these divisions have been artificially contrived, the desire to establish not just another account, but an encompassing framework as a basis for relating differing models, and the construction of a sufficiently general model of inductive discovery.

In particular, it is intended to show in this thesis that the varieties of induction and discovery all involve essentially the same kind of reasoning, but with that reasoning being controlled and distinguished through different motivations and priorities on the part of the reasoning agent. The contributions of this thesis can be stated as follows:

- The development of an encompassing framework that includes all stages of inductive discovery. This will provide a basis for evaluating and comparing different models and a means for integrating the various component parts. The framework should

include not just those stages (see Chapter 2) which are immediately obvious and lend themselves easily to computational and psychological models, but also those stages which are difficult to address and often ignored because of the limitations of current technology and research, and the problems of integration.

- The development of a model of inductive reasoning using this framework based on the scientific discovery paradigm. Basing the framework on a particular paradigm provides a frame of reference for discussion and debate of the different elements. The scientific discovery paradigm is a view of induction that we take to be useful and effective because of the emphasis on a methodology and procedures for reasoning which allow wide and easy application.

- The extension of the scientific discovery paradigm of induction to apply both to scientific and non-scientific domains. Most AI (as opposed to psychological) research on discovery has concentrated on purely scientific domains, much of it with assumptions of idealized data that are often associated with science. Mechanisms of scientific discovery and reasoning should also be capable of use in non-scientific domains which more readily admit a less idealized model of the world.

- The extension of the model of induction to consider the subjective factors such as goals and motivations that are necessary for a complete account. Real world problem solving in both scientific and non-scientific domains involves both objective and subjective elements. The richness of scientific reasoning is due to the guidance of a basic mechanism by the more varied and subtle influences of subjective collective and individual factors.

- The construction of an implementation of the internal stages of the model of induction as a demonstration of its capability and effectiveness. Although an instantiation of the model as a computational implementation unavoidably loses some expressiveness for numerous reasons, it is important to demonstrate its ability, and to bring to light limitations. An implementation can be regarded as an experiment designed to test the model of induction proposed here leading to the revision and improvement of this model in a continuous process.

The research undertaken in providing this account of induction and discovery was guided by a number of operating principles which are of particular significance in terms

of its development and contribution.

- Simplicity contributes to ease of development, evaluation and refinement. The vast amount of research on AI has led to an ever growing variety of tools, and methodologies for using those tools of ever increasing complexity. Arguments for what has been called 'minimalist AI' suggest that there should be a limited range of tools and methodologies which should only be added to when they can be shown to be inadequate [86]. This is based on the premise that advances are not made by increasing the number or complexity of tools, but from a small range of simpler tools applied in useful ways. An important consequence of this approach is that it allows the merit of such simple tools and methodologies to be evaluated easily and the tools to be revised as appropriate.

- The minimalist approach to AI is also more intuitive. Simpler theories and models are far more easily understood. This thesis does not aim for cognitive validity or plausibility, but it is hoped that it may suggest avenues to explore and investigate in the development of cognitively plausible models of human reasoning. The intuitive appeal of simpler theories allows a more ready interaction with other theories and models, cognitive or otherwise.

- Theoretical frameworks and models should not be tied to a particular discipline. The complementary disciplines of artificial intelligence, philosophy, psychology (and others) share some common goals but are subject to different traditions and emphases. Although research in a particular discipline must work to its own strengths, concerns, and abilities, it should also be accessible to other relevant fields.

- The preservation of motivations and external influences is important. Any intentional act in the world, physical or mental, is necessarily the result of the interaction of goals, motivations and other external influences. Any theory or model of reasoning must consider the role that such factors play in the larger picture.

## 1.6 Thesis Overview

More attention is being paid to the possibility and potential of automated discovery precisely because of recent progress. The significance of computer programs is also having an impact on the philosophy of science (eg. [58]), practical results being used effectively

to demonstrate certain capabilities. Much work remains, however. The next chapter discusses induction and discovery in more detail, stating more precisely how it is viewed, what it offers, and the role it has to play in reasoning. It describes a new six-stage framework for inductive discovery which encompasses *prediction, experimentation, observation, evaluation, revision and selection*, and which provides a viewpoint from which to consider related work and to identify problems and deficiencies. A brief overview of some related work is also given, providing a base for more detailed discussion subsequently.

In Chapter 3, the notion of motivation is introduced, first in general terms, and then with regard to its use in providing a control strategy for a reasoning system. A model of motivations is described and its application to different stages of discovery is discussed. Chapter 4 outlines the MID system for motivated inductive discovery. Based on the six-stage framework, a model of inductive discovery and an instantiation of that model are constructed in parallel. MID is a reasoning system that operates in the world of simple physical processes. The chapter provides an overview of the system, describing the knowledge representations, the main control strategy and the structure.

Subsequent chapters are concerned with the investigation of the individual stages of the framework. The latter stages are considered in depth, with significant details of the model and implementation being described. Chapter 5 addresses the first three stages of the framework — prediction, experimentation and observation. In the MID system, these stages are limited. The chapter discusses the role of the different stages, and considers the problems raised by each. Chapters 6, 7 and 8 address the stages of evaluation, revision and selection respectively. Finally, the results of the implementation are presented, and conclusions offered, evaluating the contribution that this work has made.

# Chapter 2

# Six-Stage Inductive Discovery

> To be able to give attention to something, it is first necessary to abstract or isolate its main features from all the infinite, fluctuating complexity of its background.
>
> — David Bohm and F. David Peat, *Science, Order and Creativity*

## 2.1 Introduction

Induction has been considered to be very many different things. This thesis is concerned with induction as a form of scientific discovery for two reasons. First, scientific discovery is a process that occurs in the real world. Many examples of actual discovery have been observed and recorded, and these provide a basis for analyses of the reasoning methods used by real scientists. This has led to the identification of temporally and physically distinct elements in the discovery process which strongly support the notion of inductive discovery as a methodology for reasoning rather than a single 'magical' process. Second, the underlying motivation behind scientific reasoning (and discovery) is one of increasing knowledge, understanding and awareness of a natural external environment in order to be able to explain, predict and possibly manipulate that environment. The second of these provides us with a large part of what we want to achieve in AI — to explain, predict and manipulate our environment. The first, if the notion of a methodology for discovery is even partly correct, provides us with a suitable means (in AI) for achieving it.

This chapter begins by discussing just what might be expected from the investigation of inductive discovery in the context of AI, and stating a position on the possibility and potential of automating discovery. Then follows the description of a framework

for a methodology of inductive reasoning which is based around notions of scientific discovery, but which subsumes other models of inductive reasoning. A brief and selective introduction to related work is then given, outlining the structure of other systems, and finding points of correspondence between them and the framework.

## 2.2 The Possibility of Automating Scientific Discovery

Exactly which elements of scientific discovery, if any, are rational or susceptible to rational enquiry, is the subject of a continued and heated debate. Views held range over the entire spectrum of opinion [58]. If we are to attempt to automate the process of discovery, however, we must be clear about what it is that we hope to achieve, and must therefore decide whether it is at all possible and if so, in precisely which parts and how.

The traditional view of scientific discovery holds that there is a clean and simple division between the contexts of discovery and justification. The context of discovery is concerned with the creation of hypotheses and theories, while that of justification is concerned with the testing of those theories and their subsequent refutation or continued use (at least temporarily). Discovery is deemed irrational and outside the scope of theories of scientific discovery, while the logical procedures of justification are capable of rational investigation (and by extension, automation). There are arguments against the rationality of justification, but these are limited and narrow, and shall not be considered here. The context of discovery is particularly problematic because it lies outside rigorous logical procedures, and is often explained by reference to insight, intuition, creativity, and a host of sociological and psychological factors. It has consequently been referred to as the 'Aha reaction'. Certainly, hypothesis formation has a richness that is due to an extensive range of experience, but to exclude it from the bounds of possibility is premature. Work on analogy, for example, focuses on just this problem in finding suitable analogical mappings for solving problems. Recent research in AI has demonstrated the effectiveness of reasoning by analogy in hypothesis formation for problem solving (see, for example, [47], [134]). Though still limited, it offers proof of the possibility of methods for discovery. Yet such methods of hypothesis formation lie outside the scope of induction, since they rely on substantial amounts of existing knowledge rather than empirical observations. Given some initial theory, however, the task of theory formation is transformed to one of theory revision of an incorrect theory based on observations. This is an altogether

different problem, and if the lack of a theory is treated as a null theory, then the theory formation problem is avoided entirely. A methodology for scientific discovery based on theory revision can as easily accommodate theories generated by other techniques (such as analogy) as it can theories revised on the basis of observation, and has the potential for the combination of such complementary techniques in a unified and integrated approach to scientific reasoning. This thesis, however, is confined to inductive discovery.

## 2.3 A Six Stage Framework for Inductive Discovery

### 2.3.1 Introduction

In response to the fragmentation of induction and discovery that has occurred over recent years as noted in the previous chapter, a new unifying framework for inductive discovery is proposed [72]. It entails six stages:

1. **Prediction.** Deductively generating predictions from a domain theory and scenario.
2. **Experimentation.** Testing the predictions (and hence the domain theory) by constructing appropriate experiments.
3. **Observation.** Observing the results of experiments.
4. **Evaluation.** Comparing and evaluating observations and predictions to determine if the domain theory has been deductively refuted.
5. **Revision.** Revising the domain theory to account for anomalies.
6. **Selection.** Choosing the best resulting revised domain theory.

The framework is a cyclical one, repeating until stability is achieved with a consistent domain theory. It begins with *prediction* which entails generating predictions for a given scenario, and then subjecting these to some kind of *experimentation.* Through *observation* and *evaluation*, the results of the experiment are compared with the predictions and, in the event that they are consistent with each other, no action is necessary. If the observations and predictions are anomalous, however, the domain theory must be *revised*, and a suitable revision *selected* to be passed through to the beginning of the cycle for use

in generating new predictions. Even when no failure occurs, the domain theory is still liable to provide anomalies at a later stage.

The framework is shown in Figure 2.1. Theories are represented by small thick-edged boxes. The original domain theory in the top left-hand corner is the input to the framework which may be a null theory if the domain is new. Shown in the figure are

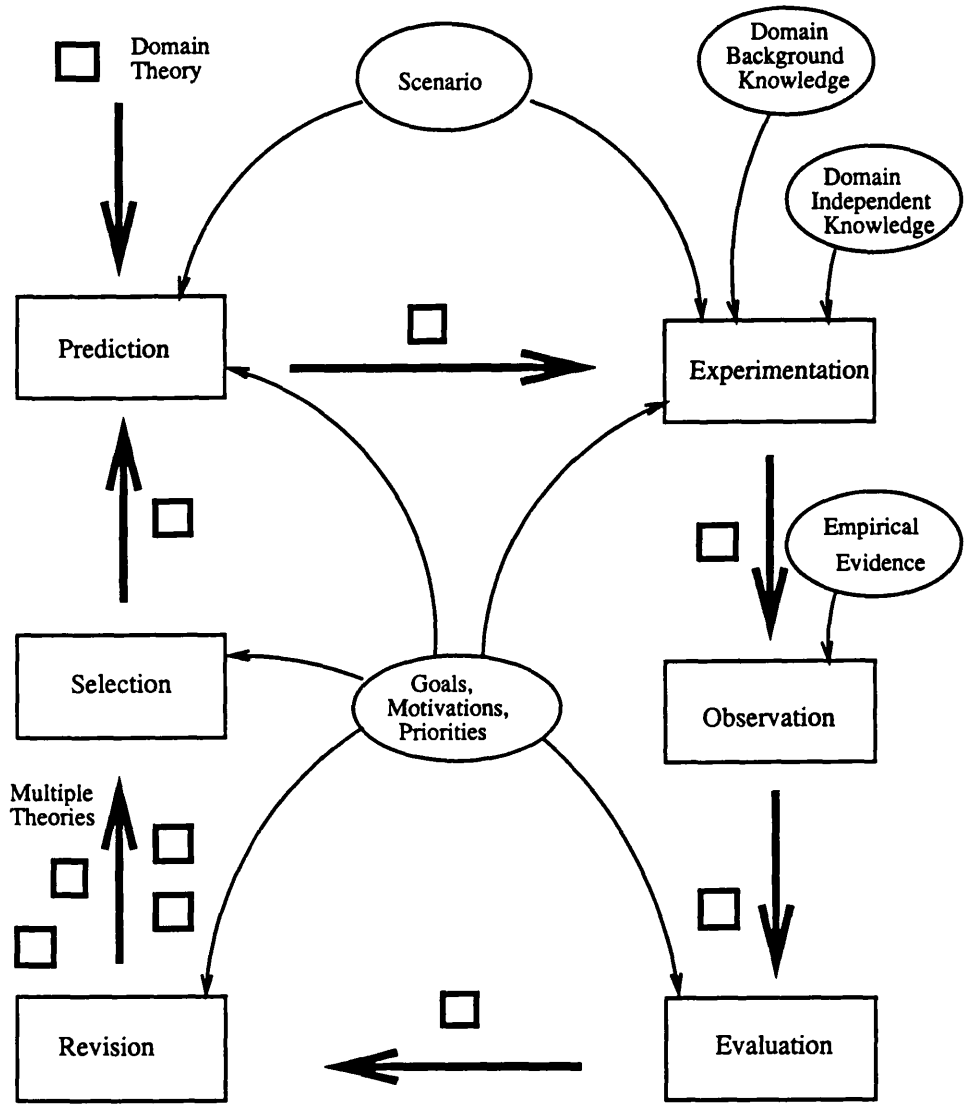



Figure 2.1: The progress of theories under the six-stage framework of inductive discovery

the different kinds of information that the framework requires in addition to the domain theory. In order to be able to design and carry out experiments, for example, substantial amounts of domain background knowledge as well as domain independent knowledge are required. Thin arrows indicate the flow of knowledge and information involved in each

stage. Thick black arrows indicate the direction of the cycle.

### 2.3.2 Prediction

Perhaps the least troublesome part of the cycle is prediction. This is a simple deductive procedure that draws logical inferences from a domain theory and background knowledge given a description of a particular scenario. In order to make sense of our environment, we continually anticipate the effects of our actions, and of external factors — we make predictions about what will happen next. Usually, our predictions are correct and we anticipate well, but there are instances when the predictions fail, and we must deal with these failures later on in the cycle.

Generating predictions can be an expensive procedure, however, demanding time and resources which may not be available. We might for example be able to predict first, second and third places in an election, yet if we are only interested in who wins, only one of the predictions needs to be generated. This is related to the motivations of the reasoning agent, in the context of which the relevance of predictions can be assessed.

It is not necessary even to have an initial domain theory here. However, if we lack a theory, then we cannot generate predictions and must experience some kind of prediction failure when we observe events not anticipated. This will lead to the gradual construction of a new theory directly from observations.

In terms of the hypothetico-deductive model, the domain theory is the hypothesis from which we draw deductive inferences which are then subjected to experimentation.

### 2.3.3 Experimentation

Once predictions have been generated, they may be empirically tested, and the results of these experiments can be compared with the predictions to determine if the domain theory (or indeed background knowledge) is, as much as possible, correct and consistent. This implies a certain requirement on domain theories that has not yet been mentioned — that they be refutable, or *falsifiable*. According to Popper [87], we may consider as scientific, only those theories which are falsifiable. Anything else, he instructs, including such diverse fields as astrology, Freudian psychology and so on, must not be considered, and must be relegated to non-science. Our position is pragmatic; in such a programme of inductive reasoning, unfalsifiable theories cannot be the subject of inference because the programme proceeds through experimentation and subsequent refutation. Moreover, an

important aim of this thesis is to show that induction and discovery do indeed apply to a broad range of domains, without regard to what is or is not scientific. Our requirement of falsifiability is necessarily independent of domain, and independent of concerns with the demarcation of science. Furthermore, in a computational implementation, we implicitly impose the restriction of falsifiability through the representation of the theory. The constraint of falsifiability constrains the kind of theory that we can reason about.

We can think of experimentation as being one of two kinds. First, there are *active* experiments in which the experimenter carefully constructs apparatus, or forces controlled environmental conditions with the aim of testing a particular characteristic or condition of a theory. Included in these are typical laboratory experiments. Alternatively, and more commonly, there are *passive* experiments which include any situation for which an expectation is generated, but for which there is no explicit theory. For example, squeezing a tube of toothpaste when brushing teeth is a passive experiment which has no controlled conditions, but which will determine if the expectation of producing toothpaste is correct or not. Both of these are important. When concerned with the problem of specifically acquiring knowledge in narrow domains, active experiments are prevalent. In normal everyday affairs, passive experiments are the norm unless they meet with a prediction failure. In this case, it is typical to switch to active experiments to find the reason for the failure, if necessary.

Thus experimentation is responsible for designing and constructing experiments in order that imperfections in the theory may be detected and corrected. This leads to observation, an important but often neglected stage in the inductive reasoning cycle.

### 2.3.4 Observation

We intend this to be a complete and encompassing framework. Were we to exclude observation, it would not be so. Although observation immediately appears transparently simple, requiring merely that changes in the environment be observed and recorded for future reference, it is a little more complicated. (It should be noted that observations may be forced by the use of controlled experiments, or may occur independently.) Observations are compared with predictions and used to decide whether the domain theory is acceptable, or whether it needs to be revised. We shall see later that recording the results of experiments is important in order to avoid oscillation in revising domain theories. In addition, there have been criteria proposed for evaluating theories, such as confirmation,

corroboration, etc., that may make use of this observational data.

Ideally, we would want an independent observer, a system capable of perceiving the external world, filtering out irrelevant information, and providing observations as input to the reasoning system. This is some way away. Even if it was possible to provide such an observer, there are definite difficulties, and some suggest that observation cannot be objective and can only be possible in the context of some existing domain theory. In other words, it is suggested that observations are interpreted before they enter the reasoning system[1]. For the moment, this is irrelevant since the point at which we can construct such a system has not yet arrived, and it is beyond the scope of the current research. Nevertheless, an appreciation of the difficulties ahead is important to this framework.

### 2.3.5 Evaluation

At this point, the experiment has been carried out, the observations have been recorded, but it remains to decide whether or not the domain theory has been falsified, whether or not it is acceptable. To make this decision, we need to be aware of a number of influential factors and to evaluate the evidence in this light. Principally, this is concerned with the quality of the evidence. If an inductive reasoning system is to be of value, then it must be able to cope with both experimental and observational error, and must be able to evaluate them in an appropriate context. Little needs to be said about the occurrence of errors, for it is undeniable that they are always present to some degree. It is, however, unacceptable to pretend to cope with them by introducing simple tolerance levels. Experimental evidence must be evaluated relative to the current motivations of a system, taking into account the implications of success or failure. In medical domains, for example, even a small degree of error may be unacceptable if it would lead to the loss of a patient's life, while weather prediction systems may, in certain circumstances, allow a far greater error tolerance.

### 2.3.6 Revision

If it is decided that the domain theory has been falsified, then it must be revised so that it is consistent with the falsifying observations. Alternatively, new theories may be introduced or generated by another reasoning technique such as analogy, case-based

[1]This is a contentious issue, and the subject of much debate. Hacking [34], for example, argues against this.

reasoning, etc. The problem of creating new theories beyond direct observation is outside of this framework. Yet we do allow for their introduction into the inductive cycle, and in addition we allow for new theories based solely upon direct observation.

Revisions to the domain theory should include all those possible within the restrictions of the knowledge representation used that are consistent with the observations. This leads to the problem of combinatorial explosion, however, and the revision process should therefore be additionally constrained by heuristic search, the search heuristics being considered in the next and final stage. Allowing all revisions, potentially at least, is important in order that they are not pre-judged out of context.

### 2.3.7 Selection

As mentioned above, this is not really a separate stage, and proceeds in tandem with revision, but the task is distinct. Since the number of possible revisions to a given domain theory is extremely large, there must be criteria for selecting those theories which are better than others. Many criteria for rating theories have been proposed, such as simplicity, predictive power, modesty, conservatism and corroboration.

However, selection of theories must be in context. This means that the goals and motivations of a system are relevant to the task of judging which criteria are more important in evaluating a theory. The way in which these criteria are applied depends upon the context in which they are used and the need for which they are used. For appropriateness of use in many situations, we may prefer Newton's laws to Einstein's, but in other circumstances, only Einstein's may be acceptable.

### 2.3.8 Summary

In these six stages lies our framework for inductive reasoning. We reflect Lakatos' method of proof and refutation [57], proposing, refuting and revising theories as necessary and appropriate until we arrive at a theory which suffices for the particular purpose at hand. More than that, we see this as a continuing process, always waiting to be invoked at the next inconsistency which is unlikely to be far away.

It should be pointed out that the temporal ordering on stages is not strict, and that a degree of interaction between stages is possible and sometimes necessary as will be discussed later. Briefly, though, there are three main times when this occurs. Prediction and experimentation are intimately related, since predictions are made in the context

of some situation or experiment. Evaluation and observation (and to some degree experimentation) are also linked in that evidence judged to be inadequate may require re-observation (or re-experimentation). Finally, part of the selection stage occurs in tandem with revision, constraining the space of revisions that can be generated.

## 2.4 Related Work

There has been, over recent years, a dramatic increase in the amount of research concentrating on aspects of discovery. In general, although many systems have been developed, little effort has been made to develop domain and implementation independent, general frameworks in which particular models or implementations can be viewed. The six stages proposed here identify those elements that are necessary for an effective system for inductive discovery. It is not the intention of this thesis to give yet another general review of existing systems. In later chapters, however, related work will be drawn on to justify and compare with this research. Below, therefore, a brief introduction to various systems is given, primarily intending to show the diversity of structure and relation to the six stage framework. It is not intended to be complete, and other systems will be discussed in other chapters as appropriate. Nevertheless, those considered here span a wide range, covering numerical (or quantitative) discovery (BACON), qualitative discovery (COAST), integrated discovery (HDD and STERN), historical discovery (KEKADA), and psychological discovery (SDDS). We begin with GRI which is used mainly to introduce the notion of dual search spaces, used by a number of other systems below.

### 2.4.1 The General Rule Inducer

An early attempt at unifying diverse approaches was Simon and Lea's General Rule Induction (GRI) program [110] which brought together problem-solving and concept formation (or rule induction) tasks. Both are information-gathering tasks, and employ guided search processes. The difference between the two is that rule induction requires search in two problem spaces — a space of rules or patterns and a space of instances or data — while problem solving requires just one — a space of rules.

In normal problem solving, the goal state is known in advance. New states are generated through a search of the rule space, and these are tested by checking to see if the goal state is a member of these. Since the goal state is part of the rule space, a

second space is unnecessary. In rule induction, on the other hand, no goal state is known ahead of time. Hypothesized rules cannot be tested directly, but only by applying them to instances, and then checking to see whether these applications give the correct result. These instances form a separate space complementary to the rule space. If the two spaces are connected, however, information from each can be used to guide the search in the other, allowing mutual heuristic search.

### 2.4.2 KEKADA

The KEKADA system described by Kulkarni and Simon [53, 54, 55] is a simulation of historical discovery. It models Krebs' discovery of the urea cycle, and draws on detailed analyses of the actual manner in which the work was carried out. The system is based on the two-space model of learning with an experiment space and a rule space. KEKADA is a production system which uses sixty-four heuristics divided into roughly equal groups of domain specific and domain independent productions. There are nine classes of production which are the basic components of the system, the first two below used for search in the experiment space, the others in the hypothesis space:

**Experiment-proposers** propose experiments.

**Experimenters** carry out experiments.

**Hypothesis or strategy proposers** decide which hypothesis or strategy to focus on.

**Problem-generators** propose new problems for the focus of attention.

**Problem-choosers** choose the next task to be tackled.

**Expectation-setters** determine expected results.

**Hypothesis-generators** generate new hypotheses about unknown phenomena.

**Hypothesis-modifiers** modify existing hypotheses.

**Confidence-modifiers** modify confidences in hypotheses based on experimental results.

KEKADA effectively simulates the discovery of the urea cycle, including the pursuit of unproductive paths on the way. The concentration on the single historical episode of discovery undoubtedly limits the system, yet it is a significant contribution to the wider field. In relation to the framework proposed here, we can group the heuristics as

follows: prediction in expectation setters; experimentation in experiment proposers and experimenters; evaluation in confidence-modifiers; revision in hypothesis-generators and hypothesis-modifiers; and selection in *decision-makers* which are used by hypothesis or strategy proposers.

### 2.4.3 SDDS

Klahr and Dunbar in extending GRI, view scientific discovery as dual search (SDDS) through a space of hypotheses and a space of experiments [48]. They carried out experiments simulating scientific discovery (using a programmable vehicle) in which subjects were required to discover new functions as program commands for the vehicle. Results led to the identification of two groups of subjects with distinct strategies: *theorists* who proposed theories and then tested them; and *experimenters* who carried out experiments and used the results to infer theories.

Based on their findings, Klahr and Dunbar constructed a model comprising three main components.

**Search hypothesis space.** This generates a fully specified hypothesis which may then be used in the next stage.

**Test Hypothesis.** In order to test the hypothesis, an appropriate experiment is generated, a prediction made, and the results observed. This produces a description of evidence for or against the current hypothesis.

**Evaluate Evidence.** The cumulative evidence is evaluated to determine whether the hypothesis should be accepted or rejected.

Although the model is quite detailed at a number of lower levels, it was not implemented in a computer program, but was intended as a specification of the control structure for one yet to be built. At this highest level, these components exclude many stages of the six-stage framework, but at lower levels some are revealed. Prediction and observation are subprocesses of experimentation (test hypothesis). Revision and selection can be taken together to be equivalent to the search of the hypothesis space, but they are not explicitly identified.

### 2.4.4 HDD

Reimann [99] investigates scientific discovery learning processes in the context of experiments with refraction. He uses an analysis of experiments with novice human subjects attempting to learn about refraction as a basis for developing a program (with a series of extensions) to model these processes. HDD, the *Hypothesis Driven Discoverer*, is formulated as an extension to and in terms of GRI which views discovery as a search in two problem spaces, one for experiments and one for hypotheses. It is intended not as a simulation of any particular subject, but as an abstract prototype learner which is effective at problem solving for the task at hand. The program is based on a production system shell with rules having condition parts on the right-hand side, and equation parts on the left-hand side.

The task is to find quantitative rules which characterize the relationship between angles and distances of objects and light rays so that the direction of refracted rays may be predicted. It is said to be a problem of *descriptive generalization* or *function induction*. Since the problem in HDD involves the incremental introduction of instances and does not have all the data available immediately, the generalizations must be augmented with other processes for modifying them in the event of inconsistencies. These include *condition induction* for modifying the condition part of the rules. More general rules are generated first so that only discrimination (specialization) is necessary in modifying rules. Differences between HDD and GRI include the induction of equations rather than rules, the attachment of conditions to these equations, the selection of appropriate attributes (in determining which features of an experiment are relevant), the use of multivalued feedback, and the construction of experiments. In actuality, HDD does not address some of these issues.

Reimann provides a model description for HDD which involves five steps:

**Step 1 Designing an experiment.** An experiment design is provided to the system.

**Step 2 Making a prediction.** One prediction is derived from applicable hypotheses.

**Step 3 Evaluating the prediction.** The prediction is compared with the actual result (the ray path) provided to the system, and either a description of the difference between prediction and result, or a statement that no difference was found is produced. No distinction is made between *approximately correct* predictions and *wrong* predictions.

**Step 4 Evaluating and modifying the hypothesis.** If a hypothesis is wrong, a discrimination process is triggered to attach new conditions to it using information about the failure so that it is corrected.

**Step 5 Generating new hypotheses.** If the current hypothesis is incorrect (resulted in a wrong prediction), then new hypotheses (rules) are created through trend-detection and function induction.

The breakdown of the model into stages shows a strong correlation with our framework. Experimentation and prediction almost directly correspond to steps 1 and 2. Evaluation of evidence is identified in step 3, but ignores important aspects. Steps 4 and 5 both deal with revision, but in different ways, depending on the kind of failure.

### 2.4.5 BACON

BACON, developed by Langley et al. [59], [61], [60], is really a suite of programs, most of which are strongly related. The BACON system searches for regularities in data in an effort to discover numeric laws. It is based around three main processes:

**Gathering data.** Given a set of dependent and independent variables, BACON organizes the data by varying appropriate independent variables and recording the values supplied by the user.

**Discovering regularities.** From the data supplied, BACON looks for constant, linear, and monotonically increasing and decreasing relations between variables.

**Defining terms and computing values.** Once BACON has found a relation between variables, and depending on the relation found, it forms new terms and computes new values for them from existing terms. This is designed to produce new terms which have constant values.

Among the accomplishments claimed for BACON, are the discovery of Boyle's Law, the Law of Universal Acceleration, Ohm's Law and Kepler's Third Law.

The search through the data space is exhaustive, and all values are supplied by the programmer. In the different versions of BACON, the search through the law space is different. The initial version searches through the data space, instantiating all independent and dependent variables, and only when all of the data has been gathered does it

search through the space of laws, looking for constant or linear relations, or defining new terms when discovering non-linear increasing or decreasing relations and then searching further. Other versions search for laws as the search through the data space progresses, using the laws discovered at lower levels as data for the laws at higher levels, and using the search through the data to guide the search through the space of laws.

BACON, as recognised by its creators, is a very simple and restricted system. It is capable of noting regularities in data at a number of levels, and thus 'discovering' scientific laws. Observation comes under data-gathering, but the system is closed to external influences and all raw data is explicitly provided. There is a limited role for evaluation (and selection), involving a simple check on whether values deviate from those required by more than an acceptable amount. Prediction and experimentation are implicit in the formation of equations through trend detection. Finally, revision of theories is achieved through defining new terms and postulating new relationships. In fairness, BACON's designers recognise its limitations to some degree, and advocate its use in parallel with other, complementary systems as part of a whole.

### 2.4.6 BLAGDEN

BLAGDEN is a system that is based on an architecture for theory-driven numerical discovery [112]. It models a particular episode of discovery from 18th and 19th Century chemistry. In this architecture, different levels of knowledge are used as the process of discovery progresses. Discovery begins with a weak theory which is derived from a *core theory* by instantiation or specialization. Sometimes, a weak theory already exists and this can be used to generate a more specific weak theory. Core theories are theories that are assumed to be true and are retained when a succession of related theories are developed and refined. Weak theories describe the relationship between variables that allow predictions to be made. They specify the factors involved, and the type of function, but do not allow precise calculations.

This weak theory is then adapted to account for the current situation by proposing informal qualitative models (IQM) which provide structural descriptions of the situation. The degree of precision and completeness of an IQM depends on the nature of the domain and the amount of knowledge available, and several IQMs may be compatible with a single weak theory. These are then used to construct *law frameworks* which specify the set of independent variables and describe the function relating them to the dependent

variable. They delimit the space of laws. After designing experiments, actual input data is used to infer *predictive laws* in the final step. Throughout the process, different kinds of knowledge are used, including the core theory, meta knowledge, background knowledge and heuristic knowledge.

Prediction here is implicit in the architecture as the goal of the system. Although there is no experimental component in the BLAGDEN system, the role of experimentation is acknowledged. Furthermore, the difficulties that can arise in the evaluation of evidence in real world domains are also recognised. Selection can be seen to occur throughout the discovery process as the space of theories and laws is gradually constrained until finally a predictive law is generated. The architecture provides a good and important framework for BLAGDEN, but assumes a correct initial core theory which may not always be justified.

### 2.4.7 COAST

Rajamoney proposes a system based on the *hypothetico-deductive* model for induction [91, 92]. The system is divided up into five distinct stages.

**Detection of problems** From the existing domain theory and a description of a scenario, COAST generates predictions and an explanation for those predictions. Rajamoney's first stage compares these predictions with the associated observations and in the case of a failure, moves to the body of the system.

**Hypothesis generation** This involves analysing the failure, determining the components of the domain theory that led to that failure, and hypothesising revisions to the theory to correct it. There is a finite set of possible causes for failure, and a finite set of revision operators, so that the revision is relatively straightforward.

**Experimentation-based hypothesis refutation** This comprises three parts: obtaining predictions from the revised domain theory, designing experiments to test the proposed hypotheses, and refuting hypotheses when positive results are obtained.

**Exemplar-based theory rejection** In order to ensure that previously refuted theories are not reintroduced, the system then uses a maintained history of observed phenomena to refute any remaining hypotheses not consistent with past events.

**Selection of a theory** Finally, in the event that multiple hypotheses still remain, it uses criteria of structural simplicity, simplicity of explanations, and predictive power

to select one.

Like HDD, COAST is divided up into different stages, and takes a somewhat global view of the process. The first three stages can be identified as prediction, revision and experimentation respectively. The fifth stage is explicit selection, and the fourth is a kind of selection that is characterized rather differently. COAST has been demonstrated on examples of evaporation of liquids, flow of a fluid, the dissolving of a substance in a liquid, and osmosis.

### 2.4.8 STERN

Cheng [6] characterizes scientific discovery as based around a *scientific research programme* involving the investigation of a delimited set of phenomena through theory and experimentation. He proposes a framework which serves as a basis for his STERN program, and which comprises three main aspects.

**Theory or Theoretical Knowledge** is regarded as sets of transformation functions which characterize the behaviour of phenomena or events. Three levels of abstraction for theories are identified: hypotheses, models and instances. This component also involves criteria for determining the adequacy of theoretical knowledge.

**Experiments** are considered to be 'black boxes' with input and output parameters corresponding to manipulated and controlled variables, and observations and measurements respectively. Like theory, experiments have three levels of abstraction: experimental paradigms, experimental setups and experimental tests. The reliability of experiments is considered here, too.

**Communication** between theory and experiment for information transfer makes up the final component.

STERN itself uses a number of classes of domain independent and domain specific rules. The domain independent rules include those for the following: strategy chooser, hypothesis testing, model testing, instance testing, models into hypotheses, instances into models, and tests into instances. Domain specific rules include: generate models, generate instances, compare instances and tests, generalize models, generalize instances, interpret results into instances, new paradigms, new hypotheses, experimenter. These are similar

to the heuristics used by KEKADA. Prediction is covered by generate models and instances; experimentation by experimenter; evaluation by compare; revision by generalize models and instances and new hypotheses; and selection is addressed by some of the hypothesis and model testing rules.

### 2.4.9 Summary

Each of the above systems is different, yet all possess some of the elements of the framework to a greater or lesser degree. The six stages are qualitatively distinct, yet they can be found in the above systems both combined together, and separated into a number of parts. Some of these systems provide a cleaner and stronger identification of the tasks necessary for discovery and induction than others. Some research addresses wider issues than just those included in the associated implementation (eg. HDD, STERN). All of these systems have inadequacies, however, which the six-stage framework exposes, and which will be investigated in the following chapters.

In developing a model of inductive discovery, it is important to build on solid foundations. The framework discussed here provides such a foundation upon which the model and implementation described in the next chapters are based. Over the course of these coming chapters, elements of the work outlined above will be described in more detail and placed in relation to the model and implementation.

## 2.5 Discussion

This thesis is intended not only to develop a model of inductive discovery that serves as a basis for effective reasoning spanning a variety of domains, but also to provide a solid foundation for inductive reasoning in general. Such a foundation in the form of the six-stage framework presented here, serves to identify those components of the reasoning process that are necessary in all complete models, and in so doing provides a base for development of the model of inductive discovery in subsequent chapters.

In this chapter, the six-stage framework has been described in broad terms, and each of the stages identified and outlined. A brief introduction to related work has illustrated the diversity of structure of existing systems, and has attempted to note points of correspondence between them and the framework. It thus gives a perspective on these systems which can be used to facilitate a stronger evaluation in comparison to

the model that is developed next.

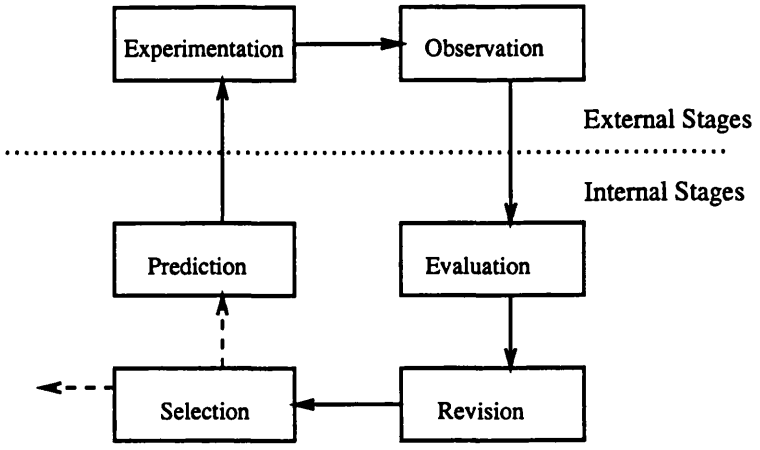



Figure 2.2: The External and Internal Stages of Inductive Discovery

The six stages of inductive discovery can be divided into two groups as shown in Figure 2.2: internal stages comprising prediction, evaluation, revision and selection; and external stages comprising experimentation and observation. This thesis concentrates on the internal stages of the framework. If experiments are to be designed and constructed, then a great deal of extra knowledge must be available. This will include knowledge of experimental procedures and apparatus, knowledge of the domain, and so on. In non-scientific domains in particular, the kind of experimentation possible may be severely limited, allowing little oportunity for directed experiments. Observation, while requiring no extra knowledge, is concerned primarily with the acquisition of evidence from an external world. In a division of the framework into internal and external stages, both of these can be considered to be external stages, while prediction, evaluation, revision and selection can be considered internal since the interaction with the outside world is minimal. In the following chapters, attention is thus focussed on developing a detailed model of the internal stages.

The model that is developed in the following chapters is in accordance with the stated aim of developing a model of inductive discovery that is not confined to scientific or otherwise restricted domains. It can be regarded as a model of qualitative discovery as opposed to numerical discovery, for example. Although it is scientific reasoning and is developed within the six-stage framework, it is not referred to as *scientific* discovery because of the restrictive sense of the word. The following chapters, therefore, describe a model of *inductive* discovery which more accurately expresses the sense of reasoning.

# Chapter 3

# Motivated Reasoning

> Knowledge for the sake of understanding, not merely to prevail, that is the essence of being.
>
> — Vannevar Bush, *Science Is Not Enough*

## 3.1 Introduction

Although there has been substantial progress in artificial intelligence over the last twenty years or so, research has tended to focus on reasoning techniques which solve problems without regard to any real external environment or to the notion of a reasoning agent. In other words, the problems and their solutions, while significant and important, have been limited in that they are divorced from real world situations. More recently, however, the importance of these issues of *situatedness* and *embodiment* has been recognised, and they are currently being actively addressed by a number of research efforts (concentrating mostly on building autonomous artificial creatures and agents, eg. [3, 2]). In considering aspects of intelligence, an agent's goals have been recognised as being important in directing reasoning mechanisms. Yet these goals have traditionally been in the framework of independently posed problems that are divorced from the problems of an agent in the real world. Motivations, which can loosely be regarded as higher level goals, provide a bridge between traditional lower level goals and reasoning mechanisms on the one hand, and real world problems facing artificial agents on the other.

Models of scientific discovery and induction to date lack a consideration of such issues, and as such have suffered from an inability to provide a fuller range of expressiveness of reasoning. Motivations provide a control strategy for reasoners, enabling such expres-

siveness through application to both scientific and non-scientific domains in a variety of contexts. This research tackles the problems of scientific induction and discovery from this perspective.

In this chapter, we consider the role of motivations in scientific reasoning. First, we introduce motivations in general and discuss the different kinds of motivation that exist. Then we consider how motivations may be modelled and represented in MID, and how they affect inductive reasoning. Finally, some related work involving motivations is described.

## 3.2 Motivated Reasoning

### 3.2.1 What are Motivations?

Research on motivation is currently being pursued from a variety of perspectives including psychology and ethology (eg. [128], [40]). Our concern, however, is with providing an effective control mechanism for governing reasoning in inductive discovery through the use of higher level goals or motivations. Though we focus on a computational approach, we will discuss related work when relevant. Some psychological research in particular, however, has recognised the role of motivations in reasoning in a similar way to that suggested here. Kunda [56] informally defines motivation to be, "any wish, desire, or preference that concerns the outcome of a given reasoning task" and suggests that motivation affects reasoning in a variety of ways including the accessing, constructing and evaluating of beliefs and evidence, and decision making. Such arguments are supported by a large body of experimental research, but no attempt is made to address the issue of how motivations might be represented.

Computational work has also recognised the role of motivations. Simon [107] takes motivation to be "that which controls attention at any given time," and explores the relation of motivation to information-processing behaviour, but from a cognitive perspective. Sloman [114, 113] has elaborated on Simon's work, showing how motivations are relevant to emotions and the development of a computational theory of mind. The current research addresses some of the issues addressed by Sloman, but in a different light. Our interest stems from a concern with discovery rather than motivation, but we recognise the relevance of motivations. Thus the ideas developed here are not aimed at cognitive modelling, but at the development of an effective computational system,

extending previous work in a new direction.

Problem solving can be considered to be the task of finding actions that achieve the current goals. Typically, goals are presented to systems without regard to the problem-solving agent so that the reasoning process is divorced from the reality of an agent in the world. Clearly, this is inadequate for research concentrating on modelling autonomous agents and creatures, which requires an understanding of how such goals are generated and selected [17]. Additionally, it is inadequate for research that aims to provide flexibility of reasoning in a variety of contexts, regardless of concerns with modelling artificial agents. Such flexibility can be achieved through the use of motivations which can lead to different results even when goals remain the same [73].

Consider the example of crossing a river. The goal is to get to the other side of the river, but the way in which that goal will be achieved depends on the motivations that generated the goal. In normal circumstances, one would look for a bridge or a boat to get across. Though this may involve more effort than swimming across immediately, it is preferable because it is more comfortable. If there are urgent reasons for crossing the river, however, such as being pursued by a wild animal, then it might be better to jump into the river and swim across instead despite the discomfort this may cause. In both cases, the motivations are different and their strengths are different, but the goal remains the same. Motivations act as a control strategy for achieving the goal, directing reasoning, and providing it with the flexibility and strength that is often lacking.

This applies equally to discovery. We might want knowledge simply to increase our understanding of the world, or we might want it so that we can achieve desired results and take actions. This is examined below.

### 3.2.2 Classes of Motivation

Much of the psychological literature stresses the distinction between two kinds of motivated reasoning phenomena (see [56] for a review). These are reasoning in which the motivation is to arrive at an *accurate* conclusion, and reasoning in which the motivation is to arrive at a particular *directed* conclusion. Kunda [56] suggests that both kinds of motivation affect reasoning by influencing the choice of beliefs and strategies applied to a given problem, but that they differ in the following respect: accuracy goals lead to the use of those beliefs and strategies that are considered most appropriate in getting the *correct* result, while directional goals lead to the use of those that are most likely to

give the *desired* though perhaps inaccurate result. According to Kunda, accuracy goals thus demand greater (cognitive) effort on reasoning, more careful attendance to relevant information, and its deeper processing with more complex reasoning strategies. Directional goals impose constraints on "search and belief construction" that lead to support for the desired conclusion.

Similar distinctions have also been noted by other researchers. In AI, Ram and Leake [98] describe two classes of goals motivating explanation at a lower level: knowledge goals which reflect an internal need for information, and goals based on accomplishing tasks in the external world. In psychology, Barsalou has distinguished between explicit problem solving goals and implicit orientation goals for maintaining a world model [63]. In education, Ng has distinguished task completion goals (such as completing an assignment) from instructional goals (what the assignment is intended to teach,) and knowledge-building goals, which relate to a student's own purposes and agenda for learning [63]. Yet another formulation of this distinction is characterized as *exploration* (knowledge, accuracy goals) versus *exploitation* (directional, task-based goals) in a number of domains. All these are mirrored in the division of motivations below into *knowledge motivations* and *action motivations*.

### Motivations for knowledge

In the context of the work here, the most important motivation is that relating to the discovery of knowledge. This can be found everywhere, even in very limited models of simple creatures, either explicitly, or by a different name such as *curiosity* (eg. [104], [77]). Any motivation which leads to the exploration of environment to discover more can be regarded as a motivation for knowledge. The desire for knowledge is relatively constant — even when action is taken to achieve some unrelated goal (to satisfy an unrelated motivation), it provides information that may be used to update a repository of knowledge. Consider, for example, eating a green banana because of hunger. Eating the banana not only satisfies the hunger motivation, but it also provides the knowledge that green bananas are not sweet. Such knowledge is always of interest and we are always motivated to acquire new knowledge even if it results from other actions. This is particularly true when considering the kind of reasoning addressed in this thesis.

### Motivations for action

Other motivations can be said to come under the broad heading of motivations for action. In this case, the motivations lead to the execution of certain actions and consequently to the manipulation of the environment in order to achieve goals. Traditional planning systems, for example, are motivated for action in that they generate plans for effecting changes in the world. These remaining motivations are thus action motivations, and include motivations such as hunger, laziness and pleasure that lead to the taking of particular actions (or exhibition of behaviour). This set is not exhaustive. Action motivations vary in strength depending on circumstances; their strength may increase to a point at which they demand satisfaction, and also decrease once they have been satisfied. In the example of crossing the river when being chased by a wild animal, the strength of the fear motivation, say, caused the immediate action of swimming across the river. After having satisfied the motivation by fleeing across the river, the relative safety might lead to the strength of the fear motivation decreasing substantially. The different kinds of action motivation are not important here. What is important is to note the difference between motivations for action and motivations for knowledge.

An example more relevant to discovery illustrating the difference between motivations for knowledge and those for action is Crick and Watson's discovery of the double helix of DNA. In attempting to become the first to discover the structure of DNA, they used 'quick and dirty' rather than the most reliable methods. Their first attempt at a model was a fiasco, according to Crick [13], partly because of "ignorance" on his part, and "misunderstanding" on Watson's. By contrast, work by Wilkes and Franklin was progressing slowly as they concentrated on using their experimental data as fully as possible, and avoided resorting to guessing the structure by trying various models. Crick states that Franklin's experimental work was first class and could not be bettered, while Watson simply wanted to get at the answer as quickly as possible by sound methods or flashy ones. While the actual motivations of the individual researchers cannot be known, their apparent motivations can be *characterized* as motivations for knowledge which demand accuracy and reliability, and motivations for action, which demand whatever behaviour will lead to the desired result.

### 3.2.3 Motivations versus Goals

It is important to establish why the use of motivations in providing a control strategy is justified. It was stated earlier that motivations can be considered to be higher level goals. If this is so, then it is not necessary to model motivations, since goals can be modelled instead. (For Sloman [113], goals are conversely considered to be motivators, but *derivative* and *nonderivative* motivators can be distinguished, corresponding to our distinction of goals and motivations respectively.) Indeed, goals are used to direct reasoning in a number of other systems. The distinction is, however, significant in the context considered here, because of the nature of the system we are considering.

In considering inductive discovery, we are concerned with finding out new knowledge, and correcting any errors in existing knowledge. However, the need for which the knowledge must be acquired is also important. Normally this need is for knowledge for its own sake in much the same way as other knowledge acquisition systems. At other times, though, there is a need for knowledge in order to take a particular action or achieve a certain result. Whether the action is brushing teeth or writing a program, the need is qualitatively different from needs with no associated actions. In both cases, however, knowledge is required and the goal of the reasoning system is the same — to acquire knowledge. The difference lies in the motivation of the system, in what motivates this goal of acquiring knowledge.

We will not address the issues of performing actions. Our goals, therefore, are always the same — they are knowledge goals, aimed at eliminating the error or inadequacy that has arisen in our knowledge. The difference between the kinds of inductive reasoning that can be undertaken is in how these goals are generated, in what motivates them. One might be motivated to take a particular action, to eat, to fight, or to do anything else, apart from simply acquiring knowledge. In all these cases, there may be a need for the knowledge that will allow the action to be taken and the desired effect to be achieved. There may also be goals generated in addition to the knowledge goals, but these lie outside the scope of the current research. Some work has been directed towards understanding how goals are generated from motivations (eg. [17]) but since we are concerned solely with the acquisition of correct and consistent knowledge, this does not demand consideration. The modelling and use of motivations is thus necessary so that the reasoning may be controlled appropriately despite the homogeneity of the knowledge-

acquisition *goals* of discovery. The relation of goals to motivations is shown in Figure 3.1. The arrows indicate the direction of goal generation. Note that action motivations are abstracted to a single class, with dashed arrows indicating goal generation links that are abstracted out. This is discussed later.

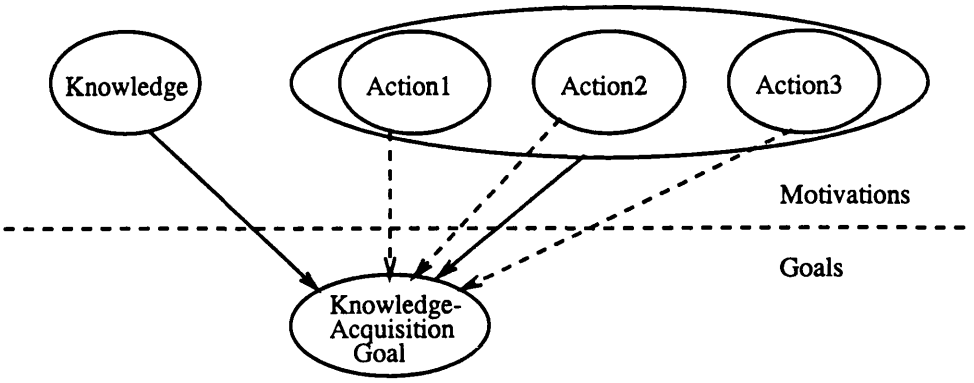



Figure 3.1: A hierarchy of motivations and goals

## 3.3 Motivation Representation

### 3.3.1 Motivation and Behaviour

Recent research into robotics, artificial life, and autonomous agents and creatures has provided the impetus for a growth of interest in modelling motivations computationally, and a number of different representations for motivations and mechanisms for manipulating them have been developed at both subsymbolic and symbolic levels (eg. by Schnepf [105], Maes [77] and Halperin [35]). This section begins with a short introduction to the modelling of motivations and continues with the development of a representation for motivations in MID.

A given stimulus does not always evoke the same response. If the external situation is constant, differences in response must be ascribed to changes in the internal state of the responding creature. These differences are said to be due to the motivations of the creature.

A creature or agent possesses a fixed range of identifiable motivations of varying strength. These motivations can be regarded as being innate, and certain behaviours may be associated with one or more motivations. For example, the behaviour of feeding is associated with the motivation of obtaining food, or hunger. In most cases, the execution

of such a behaviour reduces the strength of the associated motivations, so that in the case of feeding, the motivation to obtain food is reduced. These behaviours are known as *consummatory behaviours*; other behaviours which are not associated with any particular motivation, but which make the conditions of a consummatory behaviour come true are known as *appetitive behaviours*. For example, a go-to-food behaviour might make the conditions (that there is food within reach) of the feeding behaviour become true.

This view of motivation is somewhat simplified, and although much behaviour occurs in functional sequences with appetitive behaviours leading to consummatory ones, complex interactions between motivations and behaviours are possible [40]. For example, a single factor could directly cause many activities, or cause an action which in turn leads to other behaviours, or even cause some motivations to decrease so that others would increase in turn. In addition there are inhibitory relationships between behaviours in animals and also relationships that increase the strength of other behaviours. Moreover, the combination of motivations may lead to different or variable behaviours.

These are all difficult issues which must be addressed in attempting to construct accurate behavioural models of real and artificial creatures. Our work, however, is concerned not with behaviour, but with a prototypical reasoning agent which needs to function effectively. We can therefore ignore these more subtle issues, and concentrate on developing a simple model as appropriate.

### 3.3.2 Modelling Motivations

The specific motivations that a particular agent or creature possesses will necessarily depend on the particular kind of creature or agent being considered. Our concern here is not with modelling the motivations of any particular agent, but with providing a model of motivation that enables sufficient control to be exerted through its use over the reasoning procedures of six-stage inductive discovery. As such, we can consider a prototype autonomous reasoning agent (ARA) which we will assume can act and reason in a variety of domains and contexts. It is *autonomous* because it operates under its own motivations, and not under the direct control of someone else.

Two kinds of motivation are possible: those with constant strengths, and those with strengths that vary over time. In the creature discussed in [77], for example, the *curiosity* motivation specifically is constant, while others are variable. This constant curiosity motivation fits in well with a constant desire for knowledge that we intend for our ARA.

| *motivation* | *status* | *m-triple* |
|---|---|---|
| curiosity | constant | $< m_1, 50, \text{True} >$ |
| laziness | variable | $< m_2, 40, \text{False} >$ |
| hunger | variable | $< m_3, 197, \text{False} >$ |
| thirst | variable | $< m_4, 45, \text{False} >$ |
| aggression | variable | $< m_5, 90, \text{False} >$ |
| fear | variable | $< m_6, 80, \text{False} >$ |
| safety | constant | $< m_7, 100, \text{True} >$ |

Table 3.1: Motivation representation in MID.

The first motivation then, of the prototype ARA, is the motivation for knowledge (which encompasses such desires as curiosity, discovery), which we will fix at some constant level. The other motivations will be motivations for action, which must be specified according to the kind of agent or creature that we wish to design. These might include hunger, fear, safety, aggression, self-indulgence, and so on, and will vary over time according to the internal state of the agent. For example, if the agent spends a long time without food, then the hunger motivation will increase. When the agent feeds, the hunger motivation will decrease.

Each motivation has a strength associated with it, either variable depending on external and internal factors, or fixed at some constant value. A motivation can thus be represented by a triple, $< m, v, b >$ known as an *m-triple* where $m$ is the kind of motivation, $v$ is a real number, the strength (or intensity [113]) value associated with that motivation, and $b$ is a boolean variable taking the value *True* when the strength value, $v$, is fixed, and *False* when it is variable.

An ARA can be regarded as embodying a set of $n$ motivations, $M$, which comprises the *m-triples*, $< m_1, v, b > \ldots < m_n, v, b >$. Thus the set of motivations, $M$, is a function of the kind of agent being considered, while each motivation in this set at a particular point in time is a function of an instance of a particular kind of agent and its environment together.

Using this model, we can specify a class of agents with the motivations of curiosity, laziness, hunger, thirst, aggression, fear, safety. Table 3.1 shows an instantiation of this class as a particular agent at a particular point in time[1]. It gives the motivations, whether they are fixed or variable, and their associated m-triples. Curiosity and safety are both

[1]This agent is based on an example creature specified by Maes in [77].

fixed at strengths of 50 and 100 respectively. The other motivations all have varying strengths.

### 3.3.3 Motivations for Inductive Discovery

The model described above, and based on Maes' work, is implemented in the MID system for Motivated Inductive Discovery (described in the next chapter) in a limited way. In the case of the prototype ARA that characterizes the MID system, we might have an agent, $A$, whose motivations, $M_A$, are defined to be the set

$$\{< m_{knowledge}, k, \text{True} >, < m_2, v2, \text{False} >, \ldots, < m_n, v_n, \text{False} >\}$$

Here, $m_{knowledge}$ is the knowledge motivation (a stronger form of the above agent's curiosity) that provides the impetus for the discovery of new knowledge, and $m_2 \ldots m_n$ are the remaining action motivations which include the motivations eat, sleep, drink, safety, fear, aggression, preservation of others, self indulgence, etc. as appropriate. MID is not capable of performing actions, however, so it does not need to distinguish between the different kinds of action motivation. It does need to distinguish between the different classes of motivation that generate knowledge-acquisition goals, though. In MID, therefore, the action motivations are abstracted to a class, and MID can be regarded as having two abstract motivations: knowledge and action. This was shown in Figure 3.1. For our purpose of knowledge acquisition, these will suffice.

We will not consider how the motivations are affected by reasoning and acting in the world, but conversely how reasoning in the world is affected by the motivations. In this respect, the work here is somewhat complementary to other work that addresses these issues in the context of designing and modelling behaviour in artificial agents and creatures.

What is important in MID is to note the difference between motivations for action and motivations for knowledge, and the fact that motivations for knowledge are fixed while motivations for action can vary. This reflects a constant *desire* for knowledge, but also a recognition of the need to respond to circumstances as appropriate. Thus MID only needs to know when it is reasoning for knowledge and when it is doing so for action. Moreover, since MID does not perform actions, it has no effect on the external world. In addition, much of its information from the external world through experimentation and observation is explicitly provided. It would thus be difficult, even if included in the

model above, to show how MID's motivations would be affected by reasoning and acting in the world. For these reasons, we assume that MID is provided with a set of motivation strength values at the point of reasoning.

### 3.3.4 Dimensions of Motivation

The model of motivations described above has two principal dimensions, range and strength. The range of motivations refers to the variety of motivations that are ascribed to the ARA, and can be divided into two groups, knowledge motivations and action motivations. The knowledge motivation expresses a continual need for new information; it may be ignored if there are more pressing concerns demanding immediate attention through action motivations, but it is always present and comes to the fore in the absence of high action motivations. There is thus an important qualitative difference between knowledge and action motivations in terms of range.

The second dimension is that of strength. As mentioned earlier, each motivation has associated with it a strength value indicating its current significance. Some motivations may be fixed, especially the knowledge motivation, and these are determined by the ARA concerned. Since the strength values of other motivations are continually changing, it is important that the level at which the knowledge motivation strength, $\sigma$, is fixed is carefully chosen so that it allows an appropriate mix of reasoning for action and reasoning for knowledge. If $\sigma$ is set too high, then the reasoning will be too strongly biased towards knowledge and will not be able to react to important events; if it is too low, then reasoning will focus on action and will not allow any generally applicable knowledge to be acquired.

Figure 3.2 shows the two dimensions of motivation for the motivations described by Table 3.1. The vertical bars on the histogram correspond to each of the motivations specified, and are marked by the initial letter. In order to act or reason based on motivations, a threshold value for strength may be necessary, which must be exceeded to force action. Alternatively, the highest strength value may be used to determine the motivation currently in control. In our model we shall assume the latter, that the strongest motivation, the *salient* motivation, determines the nature of the reasoning. In the case depicted here, the hunger motivation is salient, and thus controls action. The actual way in which motivations control scientific reasoning is introduced below, and will be described in detail in later chapters which address the specifics of the different stages.

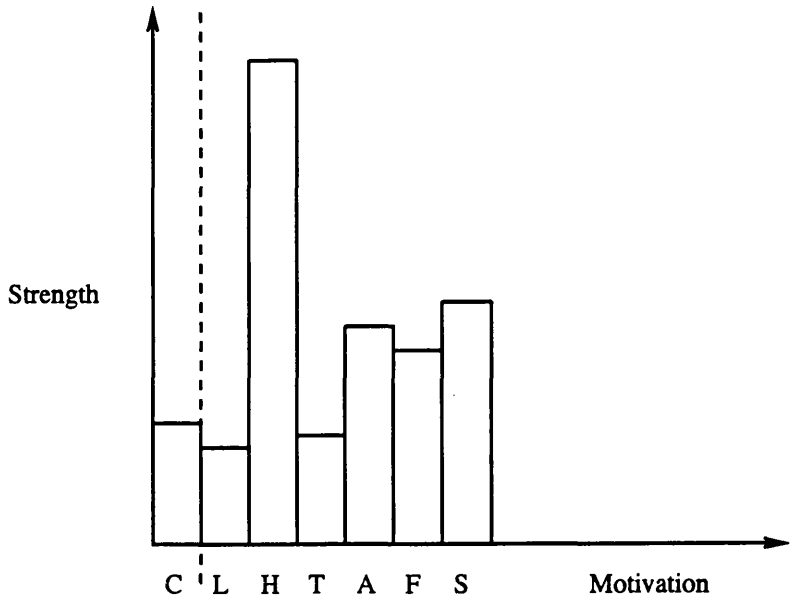



Figure 3.2: The two dimensions of motivation

## 3.4 How Motivations Affect Discovery

We have now seen how motivations may be modelled in terms of a representation and the parameters that affect the relative merits of reasoning for action and reasoning for knowledge. It still remains, however, to show how this relates to scientific reasoning, at which points in the six-stage framework it is relevant, and how so. This section provides a brief overview of the use of motivations as a control strategy in the inductive framework.

While motivation affects the design, construction and observation of experiments, we will restrict consideration to the internal parts of the framework, specifically the evaluation of evidence, and the revision and selection of theories.

### 3.4.1 Evaluation

In evaluating evidence, we want to be able to judge not only the accuracy of data, but also the validity of such data. Criticisms of a naive falsificationist approach to scientific discovery include precisely this point — that the evidence may be inadequate, insufficient or even plainly wrong. Consider, for example, testing the hypothesis that 'all swans are white', and coming across a black swan. The hypothesis can be refuted. However, it might be the case that the swan was really white and the black swan was a hallucination, or that the light was bad and what appeared to be a black swan was in fact white, or that

it was not a swan at all, but a duck. This common example is somewhat contrived, but there are real instances in which the initial scenario or observations must be modified or rejected. Karp [46], for example, notes that biologists studying attenuation often rejected evidence of the initial conditions of an experiment because of the uncertainty of their knowledge of these conditions. Indeed, we assert that the ability to recognise and reject inadequate evidence is necessary for a full account of reasoning.

The rejection or acceptance of evidence is tied intimately to the motivation that guides the reasoning process. Highly motivated reasoning implies that a successful outcome is strongly desired, and that the consequences of failure would be severe given such a high level of motivation. Clearly, in order to achieve a successful result, the evidence must be sufficiently good to enable accurate and effective reasoning, and thus the validity of the evidence used for the reasoning process acquires greater importance. Alternatively, if the reasoning is not strongly motivated, then the consequences of an inappropriate result are less severe, perhaps negligible, and less effort is demanded in evaluating evidence. Motivations determine the level of acceptability demanded of evidence according to the purposes of the ARA.

### 3.4.2 Revision and Selection

Given an inconsistency between the predictions and subsequent observations, the domain theory must be revised. There are infinitely many revised theories that can be generated that are consistent with the observations. In order to constrain the revision process, a number of selection criteria are provided that give grounds for preferring some revisions over others. However, the kinds of revisions that are preferred are dependent on the circumstances. In particular instances where the current situation demands immediate response such as in medical emergencies, it is important to generate a theory which can account for the immediate situation regardless of how well it applies to others. It may subsequently be possible to generate a more generally applicable theory, but this must be after the immediate problem situation has passed. In circumstances which are not critical, and which do not demand immediate action, more generally applicable theories can be generated at the first attempt.

The motivations of the ARA provide a measure of the kind of theory that is required. Strong knowledge motivation suggests that more effort can be spent on developing more general theories, while strong action motivations demand more specific theories which

address the immediate problem. The possibility of more variations than these are considered in Chapter 8 when discussing selection. Thus motivations determine the relative merit of the various selection criteria that are used to constrain the revision process. Existing selection mechanisms pay little if any attention to such controls, and as such are lacking in an important way. More recently, however, the role of goals and motivations in selection has been recognised (eg. [56], [98]), although much remains to be done. In Chapters 7 and 8 we show how the model of motivations specified here can be used to control the revision and selection processes.

## 3.5 Discussion

### 3.5.1 Related Work

#### Behaviour Selection through Motivation

In computationally modelling the behaviour of an artificial creature, Maes has developed a mechanism which enables both action selection [75, 76] through goals, and behaviour selection through motivations [77]. Actions and behaviours are essentially the same, while motivations are higher level goals. A creature is viewed as consisting of a set of behaviours such as feeding, sleeping, drinking, fighting, etc. Significantly, the behaviour selection mechanism relies not only on the internal motivational state of the creature but also on the external circumstances so that the creature can respond to changes in the external environment in addition to changes in motivations. At any moment, the creature is motivated towards a variety of behaviours.

Each behaviour has an associated activation level which is a real number, and a set of conditions which must be observed in order to be 'executable'. In addition, there is also a threshold which must be exceeded by the activation level when the behaviour is executable. If the threshold is passed, then the executable behaviour becomes active, and a set of processes which realize the behaviour start running. Each motivation has a strength value at a particular moment in time. The creature also has a set of sensor readings or observations which indicate its perception of the environment.

Behaviours are connected in a network with different links between them which propagate activations as appropriate amongst them. Activation energy is derived from observations of the current situation, motivation levels, and other behaviours through inhibiting

| *behaviour type* | *behaviour* | *motivation* |
|---|---|---|
| consummatory | avoid-obstacle | safety |
| consummatory | explore | curiosity |
| consummatory | fight | aggression |
| appetitive | go-to-creature | |
| consummatory | flee-from-creature | fear |
| consummatory | eat | hunger |
| appetitive | go-to-food | |
| consummatory | drink | thirst |
| appetitive | go-to-drink | |
| consummatory | sleep | laziness |

Table 3.2: The behaviours and motivations of Maes' example creature.

and exciting links (see [77] for details). At each timestep, the overall impact of the current situation and motivations is determined. If there is a behaviour which is executable, has an activation level above the threshold and has the highest such activation level then it becomes active. If none become active, then the threshold is reduced by a percentage. Once a behaviour is activated, its activation level is reset to 0. An example creature with seven motivations and ten behaviours is shown in Table 3.2.

The work has demonstrated many important points. In particular, it has shown the effectiveness of motivations as a controlling strategy in directing the behaviour of a creature, and it has provided a simple but elegant means for modelling those motivations. Underlying this work is the same basic structure as is used in MID, but it is used primarily to model behaviour in artificial creatures, and lies firmly in the robotics and autonomous agent camps. It is somewhat complementary to MID in this respect.

### Motivated Reasoning

Apart from the work discussed above, very little work has addressed the issues of motivated reasoning from a computational perspective. A notable exception is Thagard and Kunda's Motiv-PI [124], based on the PI system developed by Holland et al [41]. PI is a system for problem-solving and learning based around the use of rules. Four mechanisms underlying motivated inference were suggested:

- A representation of the self. This includes motivations and attributes (or behaviours).

- A mechanism for evaluating the relevance of a potential conclusion to the motives of the self.
- Mechanisms for motivated memory search.
- Mechanisms for adjusting the parameters of inference rules.

Motiv-PI provides a similar representation for motivations as that used here, in listing a set of specified motivations and associated strengths. The relevance of conclusions is thus determined in relation to these motivations. The motivated memory search mechanism does not merely provide a bias, but a distortion in retrieving only those instances from memory which satisfy the appropriate motivations, thus potentially generating incorrect conclusions. So too with adjusting the parameters of inference rules which is manipulated so as to generate only desirable conclusions. The system is designed from a psychological perspective, and attempts to provide a descriptive account of inference rather than a normative one, and suffers from being based around a system designed without the notion of motivation in mind. Moreover, the notion of motivation here is primarily concerned with supporting a set of beliefs about the self in terms of personality traits, and does not direct the inferencing itself. Furthermore, some elements of PI are not addressed computationally, particularly in the selection of theories, an important part of the PI system. Though limited and motivated by psychological concerns, Motiv-PI shows the potential for motivated reasoning. Many of its limitations are addressed in this thesis.

More recent research has noted the importance of motivations, but not explicitly addressed the issues. Thagard [119, 123], in response to criticisms by O'Rorke [83] proposes developing MOTIV-ECHO, a motivated version of ECHO [118], a system for theory selection. ECHO is discussed further when we consider selection in more detail.

### 3.5.2 Conclusions

The motivations of independent reasoners is of particular relevance and significance to reasoning in general. In scientific reasoning, we assert that motivations are used to control the basic reasoning strategy in order that it may apply to a large variety of domains and contexts. There is much psychological evidence to support this view [56]. (Moreover, it has specifically been proposed that motivations may have an effect on the various stages of the hypothesis-testing sequence — that is, on the generation and evaluation of hypotheses, of inference rules, and of evidence.) Using motivations to provide a control

strategy for reasoning is an approach which has not been seriously considered previously in the development of computer programs. Although motivations occupy a position at a higher level than goals, they are in a sense complementary to goals, providing a means for directing reasoning when the goals that they generate are inadequate for doing so.

Motivations thus provide a new and valuable way of controlling and directing reasoning. It is interesting to note that motivations also provide an entry point into the system for *subjective* factors. (This has also been recognised by Thagard [123].) These are subjective in that they depend entirely on the nature of the reasoning agent. One of the major criticisms of much work on scientific discovery in both philosophy of science and artificial intelligence is that it fails to take account of sociological and other factors relevant to the process of discovery [8]. Although the model of motivations developed here is very simple, there appears to be some potential for incorporating these sociological factors into the model. Furthermore, it is possible to model the motivations of a group of researchers rather than one individual. These issues clearly demand much more work and a more detailed investigation of the use of motivations in these contexts, but it seems that such work may allow the development of more complete models of scientific (and other) discovery.

This chapter has set the scene for the development of a model of motivated inductive discovery and its implementation as a computer program. It introduced the role of motivations in inductive discovery with a preliminary discussion, and continued by specifying a representation for motivations, and considering how they might be used in the evaluation, revision and selection stages of the six-stage framework.

# Chapter 4

# MID: A System for Motivated Inductive Discovery

> The only thing that counts is the ability to link this piece to other pieces ...
>
> — Georges Perec, *Life: A User's Manual*

## 4.1 Introduction

Chapter 2 described a six-stage framework for inductive discovery. Many different models can be characterized within that framework, and it therefore provides a useful means for comparisons between such models. However, although it provides a sound conceptual base from which to work, it does not provide a model in itself. Models can be considered to be instantiations of the framework, and implementations as computer programs can be considered instantiations of models. Such instantiations are important in providing an adequate account of the procedures of inductive discovery.

The remainder of this thesis is concerned with precisely that: the development of a model within the six-stage framework, and its instantiation as a computer program. We have already introduced the notion of motivated reasoning. In this chapter, we introduce the MID system for Motivated Inductive Discovery. The chapter begins with an overview of the system. Then, we discuss some of the problems with representation formalisms in discovery, and continue by specifying the knowledge representations that MID uses for its different components. Finally, we consider the relation of motivations to the rest of the system.

1. Given a domain theory, background knowledge, and a scenario description, generate predictions.
2. *Bring about the initial conditions of the scenario and also of other directed experiments to test the predictions through experiment design, construction and performance.*
3. *Observe* and record the results of experimentation.
4. Evaluate the observations with respect to
   - the adequacy of the evidence for the problem at hand through situation context and motivations of the reasoning agent,
   - and the predictions generated earlier.
5. If there are anomalies, then revise the domain theory so that it is consistent with the observations.
6. Select the best revised theory.

Table 4.1: The main control strategy in MID.

## 4.2 Main control strategy of MID

### 4.2.1 Overview

The central control strategy of MID is a simple (mostly) sequential procedure which involves each of the stages of the framework introduced and described in Chapter 2. Each of the internal stages of the framework — prediction, evaluation, revision and selection — is implemented in MID, while the external stages — experimentation and observation — are largely omitted. Nevertheless, the structure of MID acknowledges the role and position of these. Table 4.1 gives the algorithm for the main control strategy which serves as an overview of the entire system. Parts which are not implemented are italicized.

There is a one-to-one correspondence here with the six-stage framework. As noted in Chapter 2, there is some interaction between these elements, particularly with observation and evaluation when inadequate evidence may require further observation, and with revision and selection where revision is constrained by selection. This is discussed further in subsequent chapters.

### 4.2.2 Structure of MID

The main functional components of the MID program are shown in Figure 4.1. Broadly, we can divide the program into three main areas: prediction, search of the data space, and search of the hypothesis space. The dashed boxes are not implemented in MID.

MID is provided with a domain theory which encodes its current knowledge about the world — the world of simple physical processes. In normal operation, the prediction component of MID generates inferences about the future state of the world given some initial conditions describing a particular setup. Once the initial conditions are brought about by physical experimentation, the results of the experiment, as changes to objects or quantities in the world, are observed. (Predictions are used to focus attention on those changes that are expected.) Before making a comparison of the observations with the predictions, MID evaluates the adequacy of the observations with respect to its motivations (higher-level goals), and rejects them if they are unacceptable in this regard (if its confidence is too low). If the evidence is rejected, MID requests a re-observation of the results of the experiment. If the evidence is still inadequate when MID specifies the required degree of accuracy, then MID requests a new set of initial conditions (a new experiment scenario) in which the parameters of uncertainty are sufficiently good. When observations are finally accepted, they are compared with the predictions, and if there are any anomalies, then MID attempts to revise its domain theory.

In revising the theory, MID considers all possible *legal* revisions that can be made according to a set of revision operators and the constraints imposed by the observations and initial conditions. For each anomaly, MID generates an appropriate revision and ends when all of the anomalies have been resolved. However, the number of revisions that can be generated in this way is excessive, and the search space of revised theories must be constrained further by use of other criteria for theory selection. Through applying criteria on revision operators, and also criteria on states in the search space, an ordered set of revisions can be generated. Thus the 'best' revision is generated and used as the new domain theory for subsequent prediction. Finally, depending on the kind of revision, MID may also make some changes to its record of the history of events.

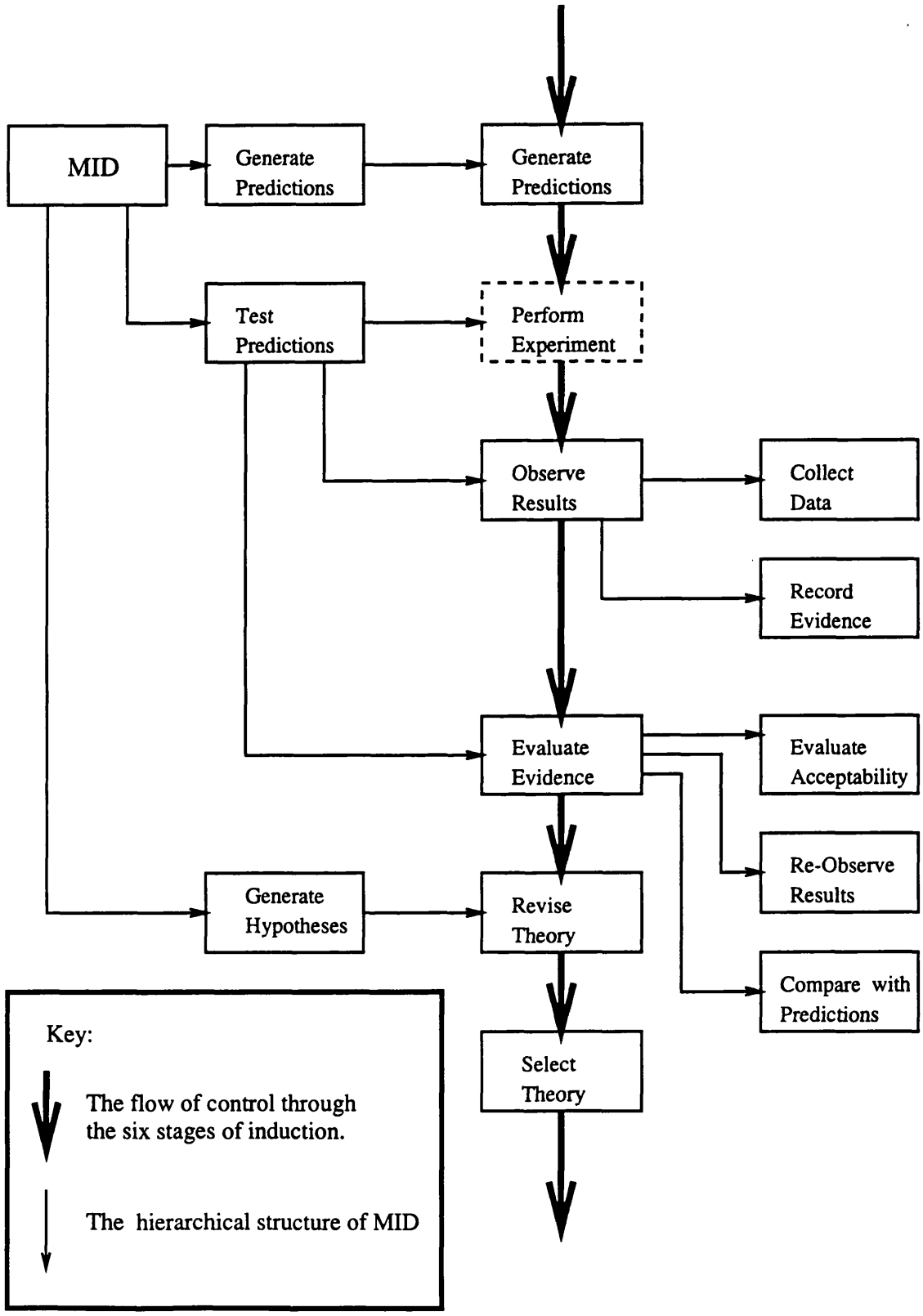



Figure 4.1: The main functional structure of the MID system

## 4.3 Knowledge Representation

### 4.3.1 Problems with Representation

Representation issues are very important. Selecting a fixed knowledge representation for use in any implementation imposes constraints which restrict the applicability and expressive power that a more general model may possess, but cannot be avoided. Different formalisms are useful for different purposes, and are appropriate with respect to these. It seems unlikely that 'unified' representations will be developed which are suited to all purposes, and we must therefore adopt a particular approach. Fixing representations does, however, have certain implications which we consider here.

In the exploration of large search spaces, for example, the constraints imposed by the representation can act as strong heuristics if the representation cannot express forms which are inadmissible anyway. The structure of a representation may also facilitate the use of simple algorithms. Roman numerals, for example, are good for addition[1]. Provided that the numbers total no more than III, the algorithm simply involves concatenation. In inductive learning tasks in particular, the representation language used is directly relevant to the space of hypotheses that can be formed. The language provides a *bias* on learning which is significant, and is increasingly being addressed (eg. [129]). More work remains to be done, especially on finding a strong and correct initial bias for representations.

In scientific discovery, the question must be raised of how an appropriate representation is developed or selected. The representation of a problem is necessarily related to the ways in which it might be solved, and the choice of representation itself is a significant part of the discovery process [11], [27], [39], [133]. Kepler for example, had to make the shift away from the representation of a law in terms of circular motion to one in terms of elliptical motion [27]. Such a shift may seem simple in retrospect, but it was a significant break with the established view of the time, and contradicted two thousand years of astronomical tradition. In addition, Cheng [10, 9] has examined the particular case of discovering the law of conservation of momentum using two different representations, mathematical sentences and diagrams. His work has shown that the discovery is 'unlikely' in the mathematical approach, but feasible under the diagrammatic approach. Though it has been claimed that finding an appropriate representation might be reduced to ordinary problem-solving tasks [108], it remains a difficult but important problem.

[1] Example given by Aaron Sloman

The issue of finding representations, however, lies beyond the scope of the research described in this thesis, and we shall adopt a particular fixed representation scheme described below, while being aware of the limitations imposed by it.

### 4.3.2 A Representation Scheme

The domain that was chosen for use in the MID system was the simple world of physical processes represented in a version of Qualitative Process Theory [22]. The domain has already been used a number of times previously in work on inductive reasoning, most notably by Rajamoney [91] and O'Rorke et al. [84], and this should allow an easier yet stronger comparison to be made between alternative models. The domain includes knowledge of processes such as absorption, evaporation, fluid-flow and solution. Although it is limited, it allows us to reason in the context of simple scientific experiments in a clear and effective manner.

Qualitative Process Theory (QPT) [22] provides a language for describing qualitative changes due to processes acting on quantities, and is well suited to such domains as this. In keeping with the principle of minimalist AI, we prefer to use a version of a widely-used and recognised knowledge representation that is suitable not just for the kind of reasoning undertaken here, but also for other purposes. The continuing research on qualitative reasoning in general and QPT in particular ensures that early redundancy of work based on it will be avoided, and that the work will be applicable in a wide variety of possible situations.

The motivation behind QPT is to represent the commonsense knowledge that people have about the physical world. According to Forbus [23], qualitative physics strives to create *wide-coverage, multi-purpose domain models* which transcend the limitations of current expert systems knowledge bases of domain and purpose-specific representations. (It is not primarily intended to model cognitive processes or representations. This is in keeping with our own aim of providing a general-purpose reasoning mechanism, though not restricting it necessarily to one which is cognitively valid.) In QPT, changes in the world are characterized as being due to *processes* of which domain theories and models are comprised. Rather than use numerical values for quantities as a basis for reasoning about change, QPT uses qualitative values such as an increase or decrease in the value of a quantity.

MID uses a simplified version of QPT, with only processes to represent distinct physi-

```
Process Name:          heat-flow
Individuals:           object ?source
                       object ?destination
                       heat-path ?path
Preconditions:         heat-connection ?source ?destination ?path
                       flow-aligned ?path
QuantityConditions:    greater-than (a (temperature ?source))
                                    (a (temperature ?destination))
Relations:             Q+ heat-flow-rate (temperature ?source)
                       Q- heat-flow-rate (temperature ?destination)
Influences:            I+ (heat ?destination) (a (heat-flow-rate))
                       I- (heat ?source) (a (heat-flow-rate))
```

Figure 4.2: A qualitative process description of heat-flow

cal changes in the world. We exclude *individual views* which are used to represent objects in a similar way to *processes* but without an *influences* slot (see below), and simply replace the *individual view* representation with a *process* representation. In fact, these are equivalent, but their different status is something that we wish to avoid because both are elements of the domain theory.

In addition to the domain theory, we have added a qualitatively different representation to encode background knowledge about the domain. This background knowledge has a different status to the domain theory, for it is accepted with a greater certainty, and contains information about the classes of objects and predicates. The Background Knowledge Rule Base (BKRB) is represented as a collection of implication rules.

### 4.3.3 Domain Theory

The domain theory consists of a collection of distinct *processes* corresponding to distinct kinds of change in the physical world. A *process* is represented as a frame which contains different slots to encode different kinds of knowledge. Specifically, the components of a frame are the *name* of the process, the *individuals* (variables) that participate in it, the conditions which are both *preconditions* and *quantity conditions*, and the effects which are divided into direct effects as *influences* and indirect effects as *relations*. Figure 4.2 shows the frame for the *heat-flow* process.

*Individuals* are objects that must exist so that a process is applicable. They are specified by an associated predicate that acts as a type constraint. Their form is as follows

individual == <type-predicate, variable>

Variables are indicated by an initial ?. A process is *active* when it satisfies the ***individuals***, ***preconditions*** and ***quantity conditions*** slots. ***Preconditions*** are predicates which specify requirements that are outside QPT in that they allow the process or activity to occur, like the requirement of a heat-connection in Figure 4.2. Their form follows.

precondition == <predicate, [variable]>

***Quantity conditions*** are requirements that are expressed within the language of the representation as quantitative relations between bodies. The general form is as follows.

quantity condition == <quantity-predicate, quantity, quantity>
quantity-predicate == <greater-than | equal-to | less-than>

In Figure 4.2, the quantity condition requires that the temperature of the source object be greater than the temperature of the destination object. The a in each of the two quantities indicates the ***amount*** of those quantities. (It is really a function from a quantity to an amount, but we will not get involved in the subtleties of QPT here.) ***Relations*** specify the indirect effects of the process through relationships between the objects on which it acts. They are of the form below.

relation == <rel-proportionality, quantity, quantity>
rel-proportionality == <Q+ | Q->

The proportionality indicates whether the two quantities are inversely or directly proportional, Q+ indicating a direct proportionality, and Q- indicating an inverse proportionality. In the example, the rate of heat flow of the process is said to increase as the temperature of the source object increases, and decrease as the temperature of the destination increases. Finally, ***influences*** specify what directly causes a quantity to change.

influence == <inf-proportionality, quantity, n>
inf-proportionality == <I+ | I->

The number n is a direct cause of the change in the quantity, and that change is either positive or negative (I+ or I-) as for relations. In Figure 4.2, the heat of the destination object is specified to increase with the rate of activity (heat flow) of the process, while the heat of the source decreases accordingly.

```
Scenario Name:   heat-flow-works-scenario
Individuals:     std-object1
                 std-object2
                 std-path1
Facts:           object std-object1
                 object std-object2
                 heat-path std-path1
                 heat-connection std-object1 std-object2 std-path1
                 heat-flow-aligned std-path1
                 greater-than (a (temperature std-object1)
                                 (a (temperature std-object2))
```

Figure 4.3: A scenario description in which heat-flow occurs

```
flow-aligned        →  aligned
fluid-flow-aligned  →  flow-aligned
heat-flow-aligned   →  flow-aligned
```

Figure 4.4: An example Background Knowledge Rule Base.

### 4.3.4 Scenarios

A *scenario* is simply a description of an external situation which is used as input to the system, encoding the initial conditions of the world, and against which the domain theory is matched to generate active processes and hence predictions. Figure 4.3 shows an example scenario in which the heat-flow process above is active. The slots are similar to those for processes. The *individuals* slot specifies the participants in the scenario, while the *facts* slot simply lists all known facts about the scenario. (In MID, facts are further divided into *q-facts* — facts with quantity predicates — and *p-facts* – the rest.)

### 4.3.5 Background Knowledge

As mentioned above, background knowledge is represented as a collection of implication rules. In MID, the background knowledge provides taxonomical and classificatory information about object types and predicates. It is held with a stronger degree of confidence than the domain theory which is the primary subject of reasoning. Background knowledge can be regarded as a set of auxiliary assumptions that facilitate the use of the domain theory. Rules in the BKRB are of the form

$$antecedent \rightarrow consequent$$

meaning that the antecedent *implies* the consequent, or that the antecedent is a subclass of the consequent. Figure 4.4 shows an example BKRB. The first rule states that

`flow-aligned` is a sub-class of `aligned`. The second and third rules state that the predicates `fluid-flow-aligned` and `heat-flow-aligned` respectively are sub-classes of `flow-aligned`. Thus we can generalize from `heat-flow-aligned` to `flow-aligned` and further to `aligned`, and specialize by traversing the rules in the opposite direction.

## 4.4 Summary

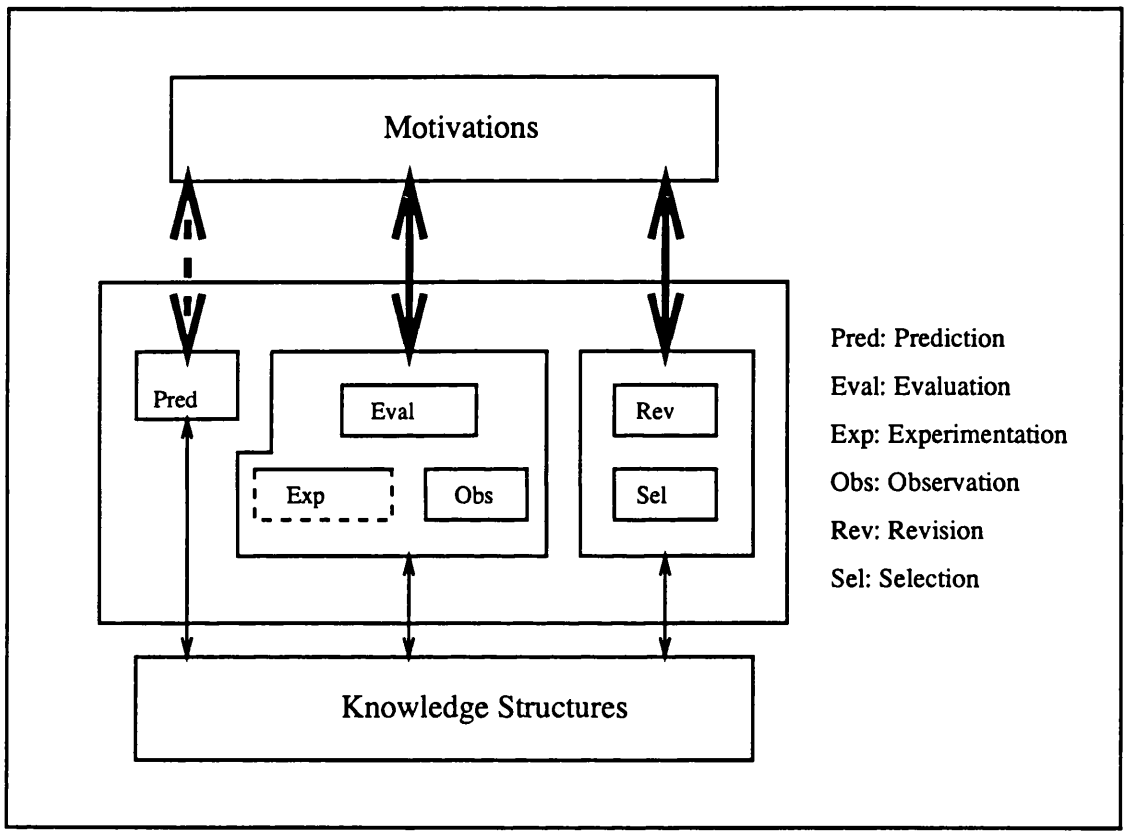



Figure 4.5: Motivations in the MID program

In the previous chapter, we discussed the representation of motivations in MID. Here we show their relation to the rest of the system. As we have seen, motivations provide an appropriate control mechanism for the reasoning process. Figure 4.5 illustrates the organisation of the MID system. At the top are the motivations of the reasoning agent, MID, which direct the application of the actual reasoning components of the system which are shown in the large box in the centre. At the bottom of the picture are the knowledge structures which MID operates on, including the domain theory, the scenario description, background knowledge and so on. Dotted lines and arrows indicate a part of the system

that has not been implemented. The thick arrows show the interaction between the motivations and the reasoning elements: motivations control the reasoning and are also affected by the outcome of that reasoning, successful or otherwise (although MID does not address the latter of these). The thin arrows indicate the interaction between the knowledge structures and the reasoning elements: the reasoning demands knowledge to operate, and the knowledge is manipulated by that reasoning. Revision and selection are grouped together because they proceed in parallel. Evaluation, experimentation and observation are grouped together because they are related, evaluation having the ability to require further experimentation and observation.

The motivations control these groups somewhat as a whole. Prediction can also be controlled by motivation; this is related to issues of relevance which MID does not address. MID assumes that all predictions that can be generated are equally relevant.

## 4.5 Discussion

As we have already noted, the representation of problems imposes constraints on the nature of the problem-solving process. A representation scheme suitable in all circumstances is not available, nor is a means for selecting between different representations dynamically as appropriate as an integral part of reasoning. Accepting that a single fixed representation scheme must be adopted, we have chosen to use a variant of Qualitative Process Theory, augmented with a background knowledge implication rule base.

All of these forms of representation are used elsewhere, and have not been specially developed for use in this research. This is significant. In keeping with our stated aim of *minimalist AI*, we avoid using special tools, preferring instead to use tools already developed, albeit with some modification. Not only does this minimize the complexity of the system, but it also makes it more widely applicable. The representations used here lie within well-defined and understood paradigms, and are supported by a solid base of existing and continuing research. There is an analogy between the methodological principles of the research described here, and what is traditionally called 'good science' in the avoidance of *ad-hoc* hypotheses. These are auxiliary hypotheses that are introduced into a theory in order to patch up a particular inadequacy with no regard to the overall plausibility of the theory. This research also seeks to avoid the introduction of *ad-hoc* hypotheses through avoiding the development of special formalisms for representation.

# Chapter 5

# Prediction, Experimentation and Observation in MID

> ...all scientific work of an experimental or exploratory character starts with some expectation about the outcome of the inquiry. This expectation one starts with, this hypothesis one formulates, provides the initiative and incentive for the inquiry and governs its actual form. It is in the light of this expectation that some observations are held relevant and others not; that some methods are chosen, others discarded; that some experiments are done rather than others.
>
> — Peter Medawar, *Is the scientific paper a fraud?*

## 5.1 Introduction

The six-stage framework described in Chapter 2 is an encompassing conceptual one designed to include the different stages of inductive discovery. The framework is important for a number of reasons. As we have already noted, it provides a sound base from which to investigate the elements of inductive discovery in more detail. Moreover, within the structure of that framework, we can develop a model for inductive discovery and construct an implementation. There are, however, stages in the framework which are not modelled in the same detail as the other stages, nor are they implemented in the MID program. As we noted in Chapter 2, the stages of experimentation and observation are *external* stages, and are not addressed in the same way as the *internal* stages. Providing a framework allows inductive discovery to be viewed as a complete reasoning methodology that does not arbitrarily exclude these external stages. Nevertheless, there are elements

of experimentation and observation that are not considered in detail here because of their external nature.

In this chapter, we consider the first three stages of the framework — prediction, experimentation and observation — discussing them in general terms at first, and subsequently in relation to the MID program. It is worth reiterating the point here that our concern is with computational discovery, and though some philosophical problems are mentioned, they are not of direct interest in themselves. We begin with prediction, discussing its significance, and then how it is implemented in MID. Then we consider experimentation, outlining some of the difficulties and discussing what can be achieved. We describe the very limited way in which experimentation manifests itself in MID, and finally consider related work and what it has to offer. The last section addresses observation, again beginning with a discussion of some of the problems facing attempts at automation, and continuing with a description of observation in MID.

## 5.2 Prediction

Prediction is perhaps the most important stage in the framework. A significant goal is to be able to mitigate the effects of our environment, and that can only be done if we are able to predict what will happen in it. Moreover, in our attempt to increase and improve our knowledge on the basis of inductive discovery, we need predictions to test against observations in order that inadequacies and inconsistencies can be exposed. The fundamental purpose of science according to Ziman [137] is to acquire the means for reliable prediction, and that is achieved by using prediction itself.

Note that prediction is similar to some concepts of explanation. This is sometimes considered to be the mechanism underlying discovery and science, but this implies that the results of the experiment are known. In the view of some, explanation consists of providing a trace of the reasoning process that shows why the observations occurred. According to Shoham [106], for example, explanation produces a description of the world at some earlier time given a description of the world at a later time, whereas prediction produces a description of the world at a later time based on the description of the world at an earlier time. If no explanation can be constructed, then there is a failure of the theory, and it must be revised. Since we must justify the predictions that are generated, there is little if any computational distinction between the two. However, at a more abstract

level, the choice of prediction or explanation takes on significant meaning. Prediction implies the ability to mitigate the effects of the environment about us. It subsumes explanation, for if we are able to predict the future, then we can explain the past. Explanation, however, seems to refer to events which must have occurred previously, and while it serves to increase our understanding of the world, it implies a subservience to that world. Computationally, the prediction stage provides explanations of its inferences as any sensible system must.

The *relevance* of predictions, too, can be important. Generating predictions can be an expensive procedure, demanding time and resources which may not be available. This is related to the motivations of the reasoning agent. If motivations are to arrive at a particular *directed* conclusion, then the issue of relevance is critical, for effort may unnecessarily be spent on making redundant inferences. We might, for example, be able to predict first, second and third places in an election, yet if we are only interested in who wins, only one of the predictions is relevant, and the others should be discarded. If the agent is motivated for *accurate* conclusions, then all predictions may be relevant.

In the sense that MID is interested in detecting and correcting all errors in its domain theory, it regards all predictions as equally relevant. Relevance is thus not an issue in MID which is concerned with acquiring knowledge. (However, some form of relevance might be important if the knowledge base grows extensively, and this could be implemented by use of motivations.) Prediction in MID is discussed below.

### 5.2.1 Prediction in MID

Prediction occurs either explicitly when we try to establish what might happen as a result of effecting certain events or conditions, or implicitly when our expectations in a particular situation are not fulfilled. In both cases, we can assume some description of the situation and some knowledge about the world which leads us to believe that certain effects will result. These correspond to the scenario and domain theory described in the previous chapter. Figure 5.1 illustrates the inference engine that provides the prediction mechanism. The domain theory (augmented by the background knowledge) is matched against the scenario description in order to generate predictions.

The generation of predictions is a relatively standard procedure. In MID, the system is provided with a scenario description and a domain theory, and predictions are generated (or inferences made) based on these. (MID is also provided with background knowledge

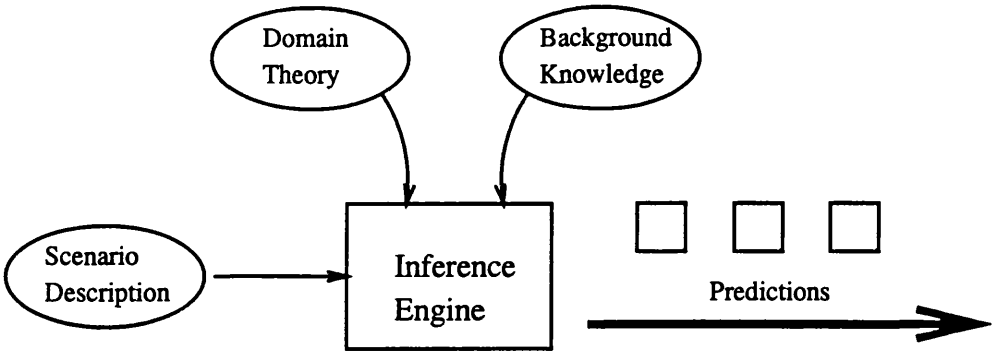



Figure 5.1: Prediction in MID

in the form of the BKRB described earlier, which augments the domain theory.) Each *process* in the domain theory is matched against the *facts* in the scenario description to determine which processes are active. First, *individuals* are matched against the *facts* in the scenario to check that the appropriate objects (quantities) exist, and to generate initial bindings for variables. In the event that *individuals* can be bound, the process is instantiated accordingly, and *preconditions* and *quantity conditions* are then matched against the *facts* of the scenario. The BKRB allows different classes of predicates in both *individuals* and *preconditions* to be used by traversing a class hierarchy along the implication rules. The process in Figure 4.1, for example, can be matched against the scenario in Figure 4.2, satisfying all of its *individual, precondition* and *quantity condition* slots. All of these conditions occur in the scenario description except for the precondition, `flow-aligned`, but this can be satisfied by using the rules of the BKRB in Figure 4.3 which state that `heat-flow-aligned` implies `flow-aligned`. If the preconditions and quantity conditions are satisfied, then the process is *active*, and the instantiated relation and influence slots (instantiated with bindings that make the process active) are advanced as predictions. All possible bindings and predictions are generated, and these are then fed through to the other stages in the framework.

In the case of the example scenario description, process (as domain theory) and BKRB given in Chapter 4, the predictions of Figure 5.2 are generated. MID generates all predictions first, and then proceeds to justify or explain each of these predictions. Here, there are two predictions, both in direct correspondence to the two *influences* of the `heat-flow` process. (The *relations* in the process specify effects which are not directly observable since they are internal to the process. Consequently, they are not counted

```
Predictions:  increase (heat std-object2)
decrease (heat std-object1)

I+ (heat std-object2) (a (heat-flow-rate))
I- (heat std-object1) (a (heat-flow-rate))

 Active:  heat-flow std-object1 std-object2 std-path1
          heat-connection std-path1 std-object1 std-object2
          flow-aligned std-path1
          greater-than (a (temperature std-object1)) (a (temperature std-object2))
```

Figure 5.2: Sample predictions generated by MID

as predictions.) The explanation consists of tracing the inference process through the instantiation of variables as `std-object1` `std-object2` `std-path1`, and satisfaction of conditions. The conditions specified in the explanation are those that must be satisfied for the process to be active, and can all be found in the *facts* of the scenario of Figure 4.2.

The output of the prediction stage is a list of the predictions that can be generated from the input knowledge comprising domain theory, background knowledge and scenario description. These predictions must then be tested against actual experimental results or observations to determine the adequacy of the domain knowledge.

### 5.2.2 Related Work

Inference mechanisms for prediction are standard. However, work by desJardins [16, 15] has addressed the utility of predictions in the context of an autonomous agent exploring and learning about its environment. Implementing a theory of Goal-Directed Learning (GDL), the PAGODA system determines which features of the world are most useful to learn. It applies a decision-theoretical technique to maximize the utility of learning goals (features) and in so doing, improves the performance of the system.

PAGODA is an important first step in considering some of the issues involved in determining the relevance of different aspects of the environment. It is, however, limited in a number of ways. In particular, utility values are fixed according to the predetermined goals of the system so that no variation is possible. Maximizing utility, therefore, will always result in the same features being learned. Although MID regards all predictions as equally relevant and attempts to correct all anomalies, it allows different motivations to determine different reasoning strategies. Further work might usefully combine the two approaches so that a representation of the internal motivations of the reasoning agent

could determine the expected utility of different features and predictions.

## 5.3 Experimentation

In a very general sense, any interaction with the outside world can be characterized as an experiment. Striking the keys of a keyboard can be considered to be an experiment testing numerous hypotheses such as that pressing the keys leads to characters being displayed on the screen, that there is a correlation between the keys pressed and the displayed characters, that the characters will be stored in a file, that the keys can be pressed, and so on[1]. In addition, there are semi-explicit experiments in which, for example, the temperature of bath water is tested with a toe before fully getting in, cartons of milk are smelled (and colour checked) before drinking, and so on. Finally, there are those which are experiments in the traditional *scientific* sense, undertaken in a controlled environment. All are experiments, and are a necessary part of the reasoning processes which govern our actions.

The term *experimentation*, however, seems to imply a directed effort at designing, constructing and performing experiments with the specific aim of falsifying or attempting to falsify theories. Indeed, it is precisely these attempts at falsification that provide the information necessary to improve and increase our knowledge about the world. Thus it is through active directed experimentation that knowledge is tested and regarded as inadequate and in need of revision. The first two kinds of experiment described above are passive experiments to a greater or lesser degree, not directed ones, and consequently require no effort of design or construction. The third kind is active and directed, and it is this kind of experiment that is considered here. There are a number of purposes to which experiments can be put: they can be used in an exploratory way to gather data for guiding the formation of hypotheses; they can be used in a directed way to test for the falsity or adequacy of a particular hypothesis; or they can be used in a discriminatory way to distinguish between multiple incompatible hypotheses.

Experimentation can be divided into the three tasks mentioned above: *design*, *construction*, and *performance*. Performance of experiments requires a physical interaction with the world. Construction of experiments requires not only physical interaction as

[1]Note that experiments are not always directed at testing particular hypotheses, and that some are purely exploratory as Hacking [34], for example, points out. He claims that it is not necessary to have an existing theory in relation to which the experiment is framed.

in performance, but also highly developed tools and abilities in order to build the appropriate apparatus. The problem of design is especially difficult, requiring extensive background and domain specific knowledge which must be applied in an appropriate way. It shares some characteristics with the discovery process, in that generating designs is similar to generating hypotheses [14].

An important difficulty with experimentation is the need for a significant amount of knowledge to be effective and competent. Without using excessive domain knowledge, there are severe limitations on the kinds of things that can be done. However, there are factors involved in experimentation over which we do have influence, and which we can control to direct the experiment design process. These include the potential to:

- determine measurements that should be made.
- suggest values for variables (eg. mass of 10g.)
- suggest variations in values (eg. increase in steps of 5g.)
- suggest quantitative changes in experiment design (eg. minimize surface area)

However, such manipulation is superficial, ignoring the significant aspects of experiment design, but still requiring substantial domain knowledge.

### 5.3.1 Experimentation in MID

The research described in this thesis focusses strongly on the internal stages of the six-stage framework. Because of the limitations described above and the external nature of the experimental stage, an experimentation component to the model and implementation have not been developed. This is not to say that experiments are not used, but that their design and construction lie outside the scope of the current research. Experimentation is necessary, and experiments are used, but they are presented to the system by the user. The *scenarios* described above provide the means for describing an experimental set-up, specifying the current world situation. Clearly, this limits the autonomy of the MID system in a significant way. However, some argue that experimentation is not always possible, and that reasoning should also proceed in its absence. Lamb [58] notes that:

> "Although it is generally advisable to test a theory wherever possible, ...it is not unscientific to adopt a theory without a test. Scientists just have to adopt theories without detailed tests. ...This necessitates that many short

cuts have to be taken, that many theories have to be taken for granted. This naturally adds to the risk that many adopted theories will be mistaken. But even the most direct knowledge claims are fallible."

Moreover, in non-scientific contexts, the empahasis on experimentation is less pronounced. We might distinguish between scientific and non-scientific discovery by considering the effort that is necessary or even possible in pursuing experimentation. Experimentation justifiably occupies an important role in science, but in more common everyday situations, the same need for designing and performing directed experiments, or the ability to do so, may not exist.

Despite some recent work on experimentation discussed below, the difficult problems of directed and controlled experimentation remain, and demand much future work.

### 5.3.2 Related Work

Rajamoney's COAST program [93, 94, 91] does provide an experimentation component. It is primarily intended as a mechanism to distinguish between multiple incompatible hypotheses or explanations through *experimentation-based hypothesis refutation.* Three strategies for experimentation are used: elaboration, discrimination and transformation. Elaboration simply entails selecting a quantity based on its ease of measurement, and refuting hypotheses which do not agree with it. All the hypotheses may, however, agree with the value of the quantity. This is not really experimentation but an evaluation of the observations with the predictions. Discrimination involves the selection of a quantity based on its ability to discriminate between hypotheses. In other words, if two hypotheses do not agree on the predicted value of a quantity, then it should be measured. Again, this is evaluation. Only the third strategy, transformation, provides any effective experimentation. If elaboration and discrimination are not sufficient to identify the correct hypothesis, then the scenario description can be transformed so that elaboration and discrimination may again be used in new circumstances. Transformation involves the modification of the scenario using a set of transformation operators that can modify the parameters of the scenario so that rates of change are different, previously satisfied conditions are no longer satisfied, or unsatisfied conditions are newly satisfied. This is very limited, and very knowledge intensive. The basic scenario must be provided to the system explicitly, and in the description of the scenario, extensive domain-dependent knowledge is required, specifying those parameters which are easily measurable, discrim-

inable, and transformable, and other specific knowledge. As a simple measure, the size of a description of a small scenario is almost trebled by the addition of the extra information needed for experimentation (see [91] for details). Furthermore, this only allows minor modifications to a basic design to be made.

Cheng's STERN program [6, 7], which simulates Galileo's reasoning strategies, also provides an experimentation component. He introduces a framework that characterizes experiments at three levels of generality: experimental paradigms, experimental setups and experimental tests. Paradigms are the most general, and involve different ways of investigating a phenomenon (eg. the inclined plane paradigm and the pendulum paradigm). Setups are instantiated paradigms with particular experimental apparatus and instruments. Tests are the most specific, and are instantiations of setups, involving precise arrangements of the particular apparatus. Each level of experiment is represented by a frame with slots for the particular characteristics and parameters of the experiment. Four strategies are identified: using experiments to (dis)confirm theories; experiment-led generalization to hypotheses; controlling the availability of experiments; and constructing new experiments. The (dis)confirmation process entails selecting an experiment, generating predictions, and comparing the results of the observations with the predictions. Alternatively, STERN uses experiments to generate results which are then generalized in a similar way to BACON. Controlling the availability of experiments is achieved through measures of ease of manufacture associated with each experiment and the number of setups. STERN can thus restrict the number of experiments according to how *practicable* it is to do so. Finally, STERN constructs new experiments if existing ones cannot generate the necessary data. This is achieved through the combination of existing experiments to produce a new integrated experiment such as a combined inclined-plane and projectile experiment. The representation of experiments includes input and output parameters which are matched when combining experiments. The combination is, however, limited in that all legal combinations are explicitly specified. STERN's abilities lie in deciding when to construct a new experiment, and then producing the appropriate frames for it. STERN also varies the values of parameters in generating experimental tests.

The IULIAN system for Exploratory Discovery developed by Oehlmann et al. [81] integrate machine discovery and case-based reasoning techniques in revising causal models by means of self-questioning, experimentation, and generation of explanations. Experimentation in IULIAN comprises three phases: design of an experiment, execution of an

experiment, and evaluation of the experimental result. Experimental design involves the utilisation of previous successful designs, and improving a design on the basis of previous design experience by using case-based planning. The experiments are stored as cases describing the problem description and the experimental result. Expectation failures and information supplied in response to questions are used to retrieve stored cases, which can be adapted in ways determined by additional questions.

Like COAST and STERN, IULIAN provides useful but definitely limited abilities. They are capable of limited manipulation of experiment descriptions which allows sufficient variation to be introduced for their particular purposes. However, much more is needed if true design and construction of useful and effective experiments is to be possible. In particular, the design of experiments in the first instance is important if we are to progress beyond the limited manipulation currently available.

## 5.4 Observation

Once experiments have been designed, constructed and carried out, the results must be observed. Naturally occurring phenomena must also be observed, as in implicit passive experiments mentioned above. On first consideration, it seems that observation is simple and straightforward, requiring merely that the appropriate events be recorded. This is naive. Paradoxically, observation is the simplest and the most difficult of the stages in the framework. At the naive level, observation simply involves waiting and watching. However, there are two kinds of objection to this view which we can call *technological* and *theoretical* objections.

Technological objections are important but straightforward. Observation of the world by a machine observer is possible, but still very limited. Although the term observation implies just vision, observation in a broad scientific sense applies to all perceptual abilities, including vision, smell, sound, and so on. The problem of integrating these different capabilities, each at an adequate level of efficiency and accuracy is a separate and difficult research area of its own, but one which has an important bearing on the possibility for independent observation as part of a complete architecture. Technological observations apply also to human observers, in the ability of the actual perceptual organs, and there are grounds for questioning the observations of such observers. Humans are also subject to limitations and deficiencies in their perceptual abilities. (What is or is not observable

is a related but difficult question [12], which will not be considered here.)

This is related to the second kind of objection, theoretical objections, so called because they refer to the role of the theory in the process of observation. Two observers of equal perceptual ability may provide different observation statements of the same phenomenon depending on their prior experience. Many examples are available where the image seen by two people is the same, but their interpretation of it is different (eg. [5]). It is argued that observation is possible only in the light of some theory, and that the observations will be expressed (and interpreted) according to that theory. Others such as Hacking [34], however, claim that theory is not necessary for observation. We will not dwell on the issue of the theory dependence of observations, but recognise that there is a contentious issue here that may not be resolved, and which causes difficulties in any discussion of observation.

### 5.4.1 Observation in MID

Like experimentation, observation is an *external* stage in the six-stage framework. In MID, observation is modelled by the explicit provision by the user of observation statements to the system as input. (Observation cannot be entirely divorced from the evaluation stage, since they are to some extent interdependent, but the distinction is clear enough for them to be treated separately.) After the stages of prediction and experimentation, the observations are provided and then compared with the generated predictions.

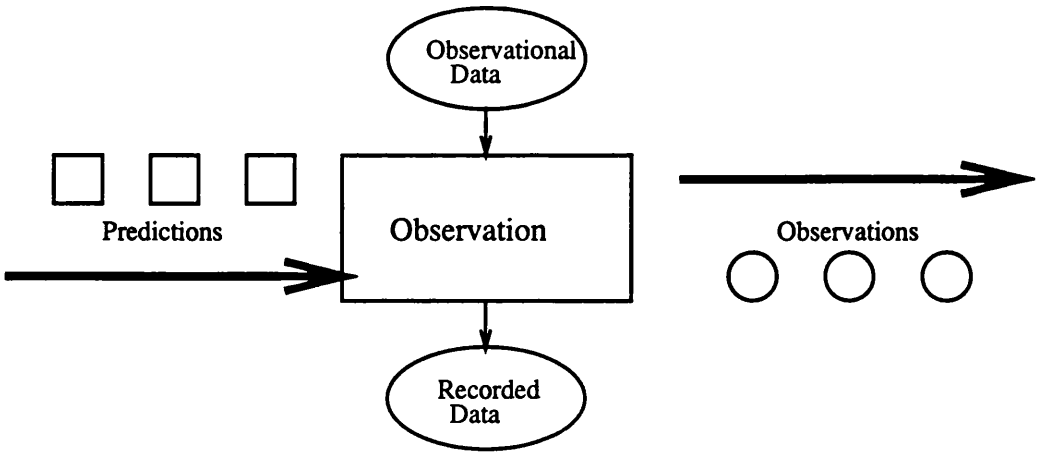



Figure 5.3: Observation in MID

In the event that there are predictions without corresponding observations (evaluated

```
Enter observations (end with return alone):
> increase (heat std-object2)
>
Is the following prediction observed?:  decrease (heat std-object1)
```

Figure 5.4: Checking observations through prediction in MID

appropriately), however, MID will attempt to re-observe by prompting the user for the appropriate observations. This allows observation to be directed in the way that it is in real world scenarios, where expectation failures often ensure a more accurate check on observations. Moreover, it introduces the notion of *relevance*, albeit in a limited and incomplete way. Although the external observer is primarily responsible for determining which events are relevant to the phenomenon under investigation, this demands that all predictions are regarded as relevant. Thus MID provides two kinds of observation: observation of naturally occurring phenomena, and observation of controlled phenomena. Figure 5.3 illustrates MID's observation mechanism.

In the example introduced earlier, two predictions were generated. Now, if only one of the corresponding events in the world is observed, say, then the user (observer) is prompted for the missing observation. This is shown in Figure 5.4.

The observation stage also entails recording the details of the observations for subsequent use in providing a measure of the evidential support of theories, and in maintaining historical consistency (see Chapter 8). In practically all other work on automated discovery, observation is implicit, and no distinct consideration of observation is typically given. This may not be a problem in the short term when the role of observation will be limited by technology. However, it fails to recognise the distinct role of observation (or data-gathering) in the real world, and thus lacks completeness. In addition, it ignores the potential of future development which may allow the integration of separate observation components into other systems.

## 5.5 Discussion

Note that use of the scenario description for prediction implies the existence of an experiment at this initial point. Particular predicted effects are generated in the context of an experiment. Thus the stages of prediction and experimentation are connected, and their ordering is not strict. In the implementation of MID, the ordering might be differ-

ent, and prediction might be delayed until the point of observation when a comparison with predictions needs to be made. The six-stage framework, however, in considering issues beyond this particular implementation, establishes a basis for this ordering. Conceptually, predictions follow from a theory, and it is those predictions that are tested by experimentation. When implicit passive experiments take place, the prediction (or expectation) exists before the experiment is performed.

The consideration here given to the first three stages of the framework has taken an external and empirical viewpoint. It should be apparent, however, that the stages of prediction, experimentation and observation are just as suited to systems where the stages are not external, nor explicit. In concept formation, for example, the prediction that each new instance presented will be accounted for by the current concept descriptions is implicit. Experimentation and observation involve the generation and presentation of the instance to the system.

In this chapter, we have discussed the first three stages of the six-stage framework in some detail, and described their instantiation in MID. Prediction and limited observation capabilities have been implemented, but experimentation has been omitted due to problems of excessive amounts of knowledge that would be required for an effective system. This concept of experimentation is, however, limited. Providing directed and controlled design and construction of experiments restricts the possibilities substantially. By leaving this stage open, many options are available, including the observation of naturally occurring phenomena and other alternatives which allow the acquisition of experimental data from different sources.

Although the focus of the research described in this thesis is not centred around experimentation, its relevance and significance is appreciated. Indeed, the six-stage framework shows the relation of the experimentation stage to the rest of the inductive programme. Work addressing the design and construction of experiments is progressing, and the framework provides a basis for the integration of this and subsequent work. Thus although experimentation is not addressed in detail here, there is an awareness of the 'bigger picture', and the current work has been undertaken with that in mind.

# Chapter 6

# Evaluation of Evidence

Our Nature of Scientific Activity teacher explained
that with scientific theories
NEAR is
sometimes close enough. ...
...Applying this idea
to what is printed here
adequacy might say
'It's there in black and white',
whereas I think the truth would rather say
two shades of grey
of which one's extremely light.

— John Hegley, *The Difference Between Truth and Adequacy*

## 6.1 Introduction

The role of the evaluation stage in the framework is essentially to determine whether or not the current theory or hypothesis under consideration is refuted by the observed evidence. At the most basic level, the decision is simple: if the evidence supports the theory, then there is no refutation; if the evidence contradicts the the theory, then the theory is refuted. In the vast majority of existing computational models of scientific discovery, this is the norm. Yet this notion of evaluation with a decision between the falsification of a theory and its continued use, is naive [5]. It ignores the possibility of error which is prevalent in most real world domains, and which can have a serious detrimental effect on the reasoning process. We identify two distinct problems: evaluation of evidence with respect to the possible error, and evaluation of theory in relation to that evidence through evaluation of prediction and observation. We refer to the evaluation of *evidence*

in order to avoid confusion with the theory selection stage (sometimes labelled evaluation) which is discussed in Chapter 8. Thus there are actually two main parts to evaluation:

- The evaluation of evidence.
- The evaluation of prediction.

There are many reasons why incorrect or inconsistent knowledge may be encoded in a domain theory. A number of alternative classifications of defective theories have been proposed by Rajamoney and DeJong [89], Mitchell et al. [80], Ellman [18], and others. There are three main kinds of problem: completeness, correctness, and intractability. A domain theory is *incomplete* if it fails to explain or predict observations. The theory might be correct as far it goes, but lacks some knowledge which results in an inability to generate the appropriate predictions. A domain theory is *incorrect* if it leads to the generation of incorrect predictions or multiple inconsistent predictions. It is *intractable* if it cannot generate predictions without consuming inordinate computational resources. The distinctions between these problems can often become blurred. In some systems, a theory missing some knowledge may generate predictions that would be considered incorrect in the presence of that knowledge. Intractability might manifest itself in the same way as incompleteness problems if no predictions can be generated. It might manifest itself as incorrectness if assumptions are used to simplify the theory. Thus we have a whole range of theory problems which are identified through comparison of predictions and observations.

We are only able to correctly identify and address these problems, however, given an appropriate evaluation of evidence. In considering the evaluation of evidence through analysis of error and uncertainty, it is clear that inadequate evidence can lead to the incorrect identification of faults in the domain theory. The need for the possibility of the rejection of evidence is thus important. It can be seen in the common practice of experimenters who often need to reject some of their data because of unreliability. In any real situation, the possibility of error in a variety of forms arises, and has important consequences for the rest of the inductive programme.

This chapter addresses the issues involved in the evaluation of evidence which determines the input to the subsequent stages of the framework. We begin by considering the kinds of error that can arise in the process of acquiring data, and establish a classification that entails four dimensions of uncertainty as a result. Next, the notion of acceptabil-

ity of evidence is introduced, and the relation of acceptance to importance is discussed. This leads to the development of a model of evidence evaluation, which is specified using the previously introduced ideas. Finally, we discuss how the evaluation of evidence is influenced by motivations, and show how motivations can control the entire evaluation procedure through a simple mechanism.

## 6.2 Error and Uncertainty

As we have seen, the six stages of induction can be split into two groups:

- External stages entailing experimentation and observation
- Internal stages entailing prediction, evaluation, revision and selection

*Experimentation* involves the manipulation of real world scenarios through the design and construction of controlled experiments in a suitable environment. *Observation*, though somewhat more passive, is also involved in interacting with the external world and serves as an entry point for evidence into the system. Any interaction with real-world scenarios must admit the inherent *uncertainty* (in a broad sense) that pervades experimentation and observation. Hon [42, 43], locates four sources of error: laying down the theoretical framework of the experiment; constructing the apparatus and making it work; taking observations or readings; and processing the recorded data and interpreting them. We have already noted some of these in the previous chapter. Problems with the theoretical framework involve background theory rather than the object domain theory. We can reduce these kinds of error from the broader classification of uncertainty as *experimental error* and *observational error* to more definite dimensions which we examine below.

### 6.2.1 Reliability

Evidence must be observed (or perceived in some way), and it must be observed (or perceived) by a *source*. *Reliability* is a measure of the ability of a particular source of evidence as the interpreter and recorder of that evidence to provide correct observation statements. Any source of observations, be it a human observer or a machine, is necessarily imperfect, resulting in some degree of variation in the *reliability* of that source. Reliability here is used in a very specific sense, referring only to the ability of a named

source in a particular situation to provide accurate observation statements of the phenomenon under investigation. The reliability of a source is in question for many reasons at different levels. A human observer, for example, may have poor eyesight which could result in the introduction of doubt into observations. In the case of a machine observer, analogous difficulties arise with considerations of the equipment used, such as the quality of lenses, and so on. The raw data of observation must also be interpreted and recorded to produce comprehensible evidence, and this conversion from sense perceptions to observation statements is another process susceptible to unreliability. Reliability is thus related to the *objective* abilities of a particular source independent of the status of the observed evidence.

### 6.2.2 Trustworthiness

*Trustworthiness* can be defined as a measure of malicious intent on the part of the information source. A source of information, even if it is absolutely reliable in terms of its ability to observe, might still provide uncertain evidence because of an intention to mislead. An independent agent acting as observer and providing input to a system in the form of observation statements has its own goals and motivations which do not necessarily correspond to those of the reasoner. Garigliano et al. [26], for example, use the case of buying a second-hand motor-bike to illustrate this. A used-bike dealer has a definite financial advantage in selling a bike and will therefore present it as favourably as possible. He might try to sell it without taking on any responsibilities but alternatively, if he is concerned for his reputation, he might give a guarantee. The evidence that he provides in attempting to sell a bike thus comes into question given his aim of getting a good deal. Buyers would only accept his evidence if confident that the dealer does not intend to cheat them. The problem of judging the dealer's *trustworthiness* is independent of problems of his reliability, since one might believe that he is fully aware of the true state of a particular bike, but is not inclined to share that knowledge. This is related to and also provides a way of dealing with the notion of *bluff* [67] which is prevalent in domains of game-playing and military strategy where agents also intentionally mislead others. Trustworthiness is thus an evaluation of the motives of an observing agent which are not necessarily the same as one's own.

### 6.2.3 Accuracy

*Accuracy* is a measure of the uncertainty arising through the evidence itself or in the method of obtaining that evidence, independent of any problems associated with an observer. In addition to the difficulties arising from the abilities and the nature of the observer, we can consider the phenomenon itself under investigation through experimentation in some form. Direct sense-perception in many instances is inadequate, especially in scientific domains, and phenomena must be observed and quantities measured using a variety of instruments and devices of varying *accuracy*. Microscopes, telescopes, rulers and other such tools are commonplace and provide the basis for observation, yet they all have limits to their accuracy. A microscope for example, has a finite degree of magnification beyond which things cannot be observed, and it also admits the possibility of inaccuracy due to faults in the lenses and so on. This kind of error occurs at a different conceptual level to that described above to illustrate the *reliability* of a machine observer. Here, the lens contributes to the *accuracy* of the data itself. As an alternative example, consider standing at a bus-stop and looking at the front of a bus as it approaches to see what number it is. While the observer may be perfectly reliable and trustworthy, the number of the bus may be partially obscured so that only the top part of it is showing. It then becomes very difficult to distinguish between a number 19 or a number 18 or a number 10, for example. Here again, the evidence that is available is itself insufficient. Accuracy is the traditional dimension of uncertainty that is considered in evaluating evidence, and is most usually addressed by the provision of simple error tolerances.

### 6.2.4 Credibility

*Credibility* is a measure of the uncertainty that arises through conflict with established constraints and prior beliefs. Although the observer may be reliable and trustworthy, and the data accurate, this does not deal with the case of exceptional circumstances when the evidence violates normal constraints. These violations include problems caused by such things as hallucinations, mirages, optical illusions and so on. What is needed is a measure of the degree of *credibility* associated with such observations. For example, the degree of credibility associated with the observation that John is flying around his office is very low (since this is more likely to be an hallucination). Alternatively, when watching a magician, the observation that a lady has been sawn in half is also not very credible. To

some extent, credibility is dependent on the expectations that are derived from existing theories, and acts as a form of conservatism. The degree of credibility will also depend on the strength with which these theories are held. Strongly held theories which conflict with observations will produce low credibility, while weak conflicting theories (including the current theory) will not affect credibility as greatly. In this respect, it reflects concerns with the theoretical framework noted earlier. Measures of credibility thus reduce the value of evidence that is highly suspect through tricks of nature.

### 6.2.5 Summary

Simple numerical measures of accuracy alone provide no information as to the circumstances in which error has originated. Our concern is with the acquisition of knowledge. As such, it is important to locate sources of error so that confidence in those sources in future situations may be adjusted in the light of the results of the current reasoning process, as well as coping with the uncertainty that currently exists. This contrasts with the use of simple error tolerances which assume a Gaussian distribution of results, an assumption which may not be justified, and which provides no useful information about the uncertainty itself.

Because of the different kinds of error and uncertainty that can arise, evidence cannot simply be accepted and used to refute or support a theory. An evaluation of observations received as evidence is necessary in order that the character of the evidence can be assessed and accepted or rejected as the situation demands. The following section addresses the issue of how to use these four dimensions to determine whether or not a theory or hypothesis is refuted by the observations.

## 6.3 Acceptable Evidence

Although we have identified four dimensions of uncertainty in evidence, the question of how reliable, trustworthy, accurate and credible evidence must be before it can be accepted and used to reason about the world remains to be answered.

On considering this issue, Levi [71] suggests a strong concept of acceptability:

> "Our beliefs guide our conduct by furnishing a criterion for distinguishing between logical possibilities which for *all* practical and theoretical purposes may be utterly ignored. In ignoring such logical possibilities, we set the risk of error involved in acting as though they were false at 0. In this sense, we are certain that they are false."

Although this is part of a larger debate concerning other issues, Levi's point is certainly valid here. The acceptance of observations as a basis for reasoning confers upon them a status beyond a mere value of likelihood or certainty; they are accepted as being true. Similarly with the rejection of observations, which to all intents and purposes are false rather than merely very unlikely or uncertain.

### 6.3.1 Confidence

The notion of the acceptance of evidence leads to a two-way split of evidence into that which is sufficiently good to be used, and that which must be rejected. Suppose that the four dimensions described above can be combined in some way to form a single measure for evidence, *confidence*. Now it can be said that if we have sufficient *confidence* in some evidence then that evidence is acceptable and may be admitted. If not, then the evidence must be rejected outright and re-observed (possibly involving a new or repeated experimental set-up). This means that there is a point of commitment to evidence beyond which observations are accepted in full and before which they are rejected outright so that the need to maintain large amounts of data on the certainty of propositions (theories and data) throughout the reasoning process is avoided.

We define $C(e)$, the confidence in a piece of evidence, $e$, to be a function on reliability, $r$, trustworthiness, $t$, accuracy, $a$ and credibility, $c$, as follows:

$$C(e) = f(r, t, a, c)$$

### 6.3.2 Acceptance Thresholds

Now, in order that a piece of evidence, $e$, may be considered acceptable, the confidence in it, $C(e)$, must exceed some limit, the *acceptance threshold*, $C_{accept}$, which marks the degree of confidence that is required for acceptance. However, this threshold value cannot be an objective quantity since the degree of confidence required in any given situation is clearly dependent on the motivations of a reasoner with reference to possible consequences of incorrect decisions based on the acceptance of faulty evidence. Consider taking a train to Scotland today for a business meeting. Since the meeting is very important, a high degree of confidence in the acquired evidence is necessary in order not to be late. This might mean rejecting the evidence provided by a co-worker, but accepting the evidence provided by the railway authorities about the times of trains. Alternatively, if there is a party rather than a business meeting, it is more reasonable to accept the evidence of a friend,

since even if that evidence is incorrect, being late for a party is not very important. These examples illustrate the relation between the importance of the situation and the degree of confidence required. Consideration of motivations can provide a simple measure of the importance of any given situation which can then be used to determine the appropriate acceptance threshold value.

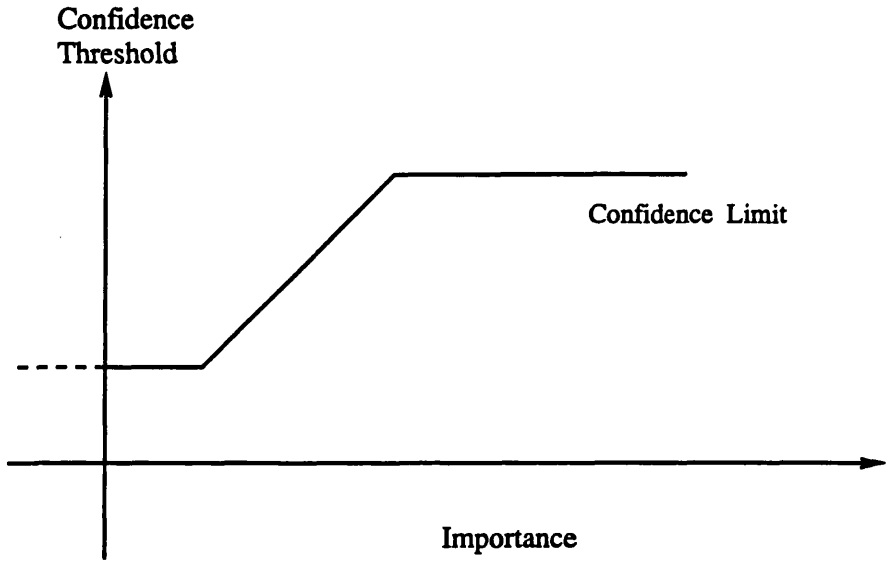



Figure 6.1: The relationship between confidence thresholds and importance

Figure 6.1 shows the relationship between importance and acceptance threshold values of confidence. With low importance, the acceptance threshold is also low since evidence with a lower degree of confidence is acceptable in such cases as in the party scenario above. As importance increases, so does the acceptance threshold in a uniform way up to a point when a confidence limit is reached where confidence increases very little if at all. Total confidence in empirical evidence is impossible due to the various ways in which uncertainty may be introduced as discussed earlier. Although confidence may be improved by using more reliable observers and tools, the possibility of error cannot be entirely eliminated. The confidence limit on the graph acknowledges some degree of uncertainty while avoiding the problem of paralysis of action. With high importance, therefore, the acceptance threshold tails off into a plateau so that the evidence can be accepted in order to continue to the other stages in the reasoning process. With low importance, there is another plateau, this time to avoid the case of reasoning based on evidence with no confidence despite the lack of importance. The points on the graph at

which the limits meet the main slope are sharp, allowing a simple model to be constructed. An alternative view could have the slope curve smoothly into the limits.

### 6.3.3 Action Points

Although the upper confidence limit above prevents a paralysis of action in situations of high importance, this fails to consider issues of *urgency*. Time is often a factor in critical situations of importance, and it may not be reasonable to reject evidence that falls short of the threshold with such time constraints. Examples might include the systems in a nuclear power plant, or in an aeroplane which has experienced some sort of failure, and so on. In these cases, importance is high, but time is also severely limited, and the confidence plateau may still be too high for the available evidence. To deal with this, there must be an *action point* which is determined by the *urgency* of the situation, and which allows reasoning to be based on evidence with a low degree of confidence — lower than the acceptance threshold for confidence. This action point is usually the same as the acceptance threshold because if there is no (or low) urgency, then there is no need to accept suitably uncertain evidence. As urgency increases and approaches a critical level, the action point decreases so that less and less confident evidence may be accepted in order to act. Note that a (high) degree of urgency implies that the reasoning is motivated for action. In fact there will always be urgency to a greater or lesser extent when reasoning to satisfy action motivations.

(The roles of importance and urgency have also been recognised by Sloman [113] in developing a computational theory of mind, but they are more thoroughly developed below.)

Whenever evidence is accepted, the manner of its acceptance is relevant to determining how it may be used. If poor evidence is accepted when reasoning for action, then it is used to provide a *local* update to the domain theory for the current situation only. It is used for *temporary* revision rather than permanent revision so that the need for action can be satisfied, but without compromising the need to maintain a correct and consistent domain theory. Further or better evidence may subsequently be obtained for use in generating a permanent revision. This is discussed in greater detail in Chapter 8 on selection.

## 6.4 A Model of Evaluation

The requirements of a system for evaluation have been discussed in general terms above. This section provides a more formal model of evaluation. First, it is necessary to specify the four parameters of uncertainty: reliability, trustworthiness, accuracy and credibility. Reliability, $r$, and trustworthiness, $t$, are defined to take values in the interval between 0 and 1.

$$r \in [0,1]$$
$$t \in [0,1]$$

The upper limit of 1 denotes perfect evidence without any possibility of error, while the lower limit of 0 denotes evidence that is perfectly incorrect. 0.5 indicates evidence which there is no reason to believe or disbelieve in terms of its reliability or trustworthiness. Thus values ranging from 0.5 to 1 indicate increasing degrees of support for the evidence, while values ranging from 0.5 to 0 indicate increasing degrees of support against the evidence. These lower values provide strong reasons for the rejection of evidence since the evidence is judged to be not merely uncertain, but misleading.

Reliability and trustworthiness are both attributes of a source of observations, and allow the possibility not only of uncertainty of their correctness, but also of uncertainty in their incorrectness. In other words, the observations provided may not merely be approximate, but deliberately misleading as discussed earlier. It is easy to see this in the case of trustworthiness, for it might be expected that a source will attempt to mislead, in which case the parameter value will be below 0.5. With reliability, however, this is less obvious, since it is a measure of the ability of a source. Consider, though, an observer who has no malicious intent (so that trustworthiness takes a high value), but who provides evidence with no basis in observation. Instances of this will be rare, but are nevertheless plausible. For example, in response to requests for information (such as the time of the last train to Scotland), people often guess answers. Mostly, there is some basis in experience for this, but sometimes there is none, or the experience may be so far removed and irrelevant that the evidence may be judged to be not so much approximate as incorrect. This is exacerbated in extreme cases, such as mental disorder, for example, where sources are trustworthy (in the sense of no malicious intent), but are unreliable to the point of providing totally false information.

Thus in evaluating *binary* evidence, very low values for trustworthiness and reliability may also provide a basis for acceptance, since the opposing evidence may be assumed

to hold. This work does not consider binary evidence, however, and the issue is not addressed further.

Accuracy, $a$, and credibility, $c$, as attributes of the phenomenon itself take values from the interval between 0 and 1.

$$a \in [0,1]$$
$$c \in [0,1]$$

The upper limit of 1 denotes perfectly accurate or credible evidence, while the lower limit of 0 denotes no degree of accuracy or credibility at all. Note that these are attributes of the phenomenon, not of its manner of observation.

Note that the meaning of the values for each of the parameters of uncertainty is relative and lies in its comparison with other values. Thus a value of 0.5 for accuracy, for example, means only that it is more accurate than a value of 0.4, and less accurate than a value of 0.6. It is possible however, to use a particular interpretation of these values, and treat them as probabilities, say, so that a value of 0.5 for accuracy would take on the meaning of a 0.5 probability of perfect accuracy. In the current work, however, no such interpretation is used. The only significance of the values is in the ordering they impose on the associated evidence.

Given values for each of the four parameters of uncertainty, we can define confidence by some function which combines them. The confidence, $C$, in a piece of evidence, $e$, is defined below, where $f$ is a function that combines positive values into a single measure for confidence. Note that if reliability or trustworthiness have values less than 0.5, then the confidence is zero. By leaving the function $f$ unspecified, different ways of combining the values of uncertainty are possible.

$$\mathrm{C(e)} = \begin{cases} 0, & \text{if } r < 0.5 \text{ or } t < 0.5 \\ f(r,t,a,c), & \text{otherwise} \end{cases}$$

Now we can define an *acceptance threshold*, $C_{accept}$, which is the amount of confidence required for acceptance of the evidence in order to be able to act on it. This is proportional to the importance of the current situation, denoted by the *importance index*, $\mathcal{I}$. However, with high importance, confidence will not increase proportionately. Similarly, with very low importance, some minimum degree of confidence is necessary. Thus we have:

$$\mathrm{C_{accept}} = \begin{cases} C_{max}, & \mathcal{I} > C_{max} \\ C_{min}, & \mathcal{I} < C_{min} \\ \mathcal{I}, & \text{otherwise} \end{cases}$$

The importance index, $\mathcal{I}$, is drawn from the interval 0 to 1. $\mathcal{I} \in [0,1]$. The upper limit of 1 denotes supreme importance, while 0 denotes no importance at all. Particular values

used derive their significance from comparison with other values of importance in the same way as for the dimensions of uncertainty. Thus a value of 0.5 indicates greater importance than a value of 0.4, but less importance than a value of 0.6. Importance is calibrated by reference to confidence and the dimensions of uncertainty. The value of the importance index is significant in the level of confidence demanded, and by extension in the degree of uncertainty permitted.

Finally, an action point can be defined by introducing the urgency $\mathcal{U}$ associated with the scenario which is also drawn from the interval 0 to 1. $\mathcal{U} \in [0,1]$. The upper limit of 1 denotes maximum urgency, and demands immediate attention, while the lower limit of 0 denotes minmum or no urgency. $\mathcal{U}_{lim}$ is a limiting value for urgency below which no reduction to the acceptance threshold is necessary.

$$\mathrm{C_{action}} = \begin{cases} C_{accept}, & \text{if } \mathcal{U} < \mathcal{U}_{lim} \\ C_{accept}.(1-\mathcal{U}), & \text{otherwise} \end{cases}$$

Acceptance thresholds and action points bear some similarity to Marsh's cooperation thresholds which also use importance in deciding when an agent should cooperate with another in order to achieve its goals [78]. However, it is not clear how measures of importance in his model are derived. In the discussion of an implementation below, the origin of importance is addressed through the use of motivations in the reasoning agent.

## 6.5 Evaluation in MID

The model presented here is implemented in MID. When evidence is supplied to the system, it is evaluated according to the specification above. If it is deemed acceptable, MID uses it as appropriate. If not, then MID needs better evidence. Details of the instantiation of the model in MID are given below. The organisation of evaluation in MID is summarized in Table 6.1.

### 6.5.1 Sources of Uncertainty

Accuracy and credibility are dimensions that are associated with a particular scenario experiment. This may be because of the instruments that are used in the scenario, or because of the nature of the scenario (a magician's stage performance, for example,) as described above. In both cases, accuracy and credibility are by definition factors of the scenario, and do not vary within that scenario. Their values are obtained with the scenario description.

Input includes domain values, $< \Delta, C_{max}, C_{min}, \mathcal{U}_{lim} >$ (where $\Delta$ is the domain theory), and a scenario, $< \mathcal{S}, a, c, \mathcal{U} >$ (where $\mathcal{S}$ is the scenario, $a$ the accuracy, $c$ the credibility, and $\mathcal{U}$ the urgency).

1. Obtain a set of observations, $< O_1, r, t >$ where $O_1$ are the observations, $r$ is the reliability of the observer and $t$ the trustworthiness of the observer.
2. Calculate the acceptance threshold, the action point and the confidence in the evidence for those observations.
3. If the evidence is acceptable, then compare the observations with the predictions and proceed to the other stages of reasoning.
4. If the confidence does not exceed the action point, then:
   - Get new observations with acceptable values for reliability and trustworthiness, $< O_2, r, t >$.
   - Calculate the acceptance threshold, the action point and the confidence in the evidence for those observations.
   - If the evidence is acceptable, then compare the observations with the predictions and proceed to the other stages of reasoning.
   - If the confidence does not exceed the action point, then request a new scenario with new values of accuracy and credibility, and go to step 1.

Table 6.1: Algorithm for evaluation of evidence

Reliability and trustworthiness differ in that they are factors of the observer, (again as described above,) and as such are associated with the observations themselves. In the real world, a whole set of observations will usually come from the same observer at once, and can therefore be evaluated at once without needing to consider individual observations. Grouping observations together strengthens the concept of an observing agent in comparison to the idea of multiple independent observations which is a less natural view of the world. An observation itself, being a record of the interaction of an observer with an uncertain world, has no special characteristics that can identify it as inadequate. It is only by evaluating the observer and the world that the resultant observations can be assessed. Thus, when we reject observations, we actually reject the observer responsible as the source of the inadequacy, and perhaps also the scenario as the environment responsible for the inadequacy.

The replacement of an observer entails new values for reliability and trustworthiness. An observer must be replaced when low values of reliability or trustworthiness contribute to an unacceptable level of confidence in the evidence. Similarly, the rejection of one

| *origin* | *parameter* |
|---|---|
| scenario | Accuracy |
| | Credibility |
| observer | Reliability |
| | Trustworthiness |
| domain | Confidence limits |
| | Urgency limit |
| motivations | Importance |
| scenario | Urgency |

Table 6.2: Parameters of evaluation and their origin.

scenario and its replacement by another involves the replacement of values of accuracy and credibility, and only when existing values cause inadequate levels of confidence.

Maximum and minimum degrees of confidence are fixed by the domain under investigation, defining the practical limits of acceptability of evidence. Theoretical absolutes cannot be attained in real world scenarios, and these delimiting degrees of confidence as parameters of a particular domain, serve to avoid paralysis of action in extreme cases. In addition, the domain also specifies a limit for urgency, $\mathcal{U}_{lim}$, above which urgency is introduced into the requirements for acceptability.

Table 6.2 summarizes the origins of uncertainty, confidence limits and motivational values. Note that the scenario is shown in the table twice; first as determining values of accuracy and credibility, and second as determining a value for urgency.

### 6.5.2 Importance and Motivation

The importance that is attached to particular circumstances is responsible for determining the required level of confidence for acceptance of evidence. Importance is a subjective factor. What is very important to one person is not as important to another in the same situation. In the train timetable example given earlier, it is clearly less important to one's colleague than to oneself to catch the train on time. At the same time, there is an independent factor which affects such considerations, namely *urgency* which can be considered as a parameter of the scenario itself.

Although variable, importance is strongly related to the goals that demand action and reasoning, and by extension, to the motivations that specify goals. Regardless of the kind of situation (the domain theory, scenario, etc.), the importance is a function

of the strength of motivation to act (or reason) in that situation. Thus the stronger the motivation, the greater the importance, and the more confidence is necessary for acceptance of evidence.

Now, as defined in Chapter 3, motivations are represented as a set of ***m-triples*** of the form $< m, v, b >$ where $m$ is the motivation, $v$ is its strength value, and $b$ is True if the strength is fixed (and False otherwise). We define the ***importance index***, $\mathcal{I}$, as follows:

$$\mathcal{I} = kv_{max}$$

where $k$ is a proportionality constant, and $v_{max}$ is the ***salient*** motivation in $M$ for the given autonomous reasoning agent or ARA (in this case, for MID). (Recall that the ***salient*** motivation is the motivation with the highest strength value. Thus $v_{max}$ is the maximum value of $v$ in $M$.) The ARA can be said to be reasoning under this motivation and, consequently, its strength is representative of the importance attached to the reasoning undertaken in order to satisfy that motivation. The stronger the motivation, the more that motivation needs to be satisfied (for unsatisfied motivations grow stronger and will subsequently demand stronger and possibly greater action) and the more important it is that the reasoning process should be successful. This requires greater effort in terms of time and resources to be spent on evaluation in order to avoid the consequences of an undesirable failure. Thus motivation determines importance, importance determines the acceptance threshold, and the acceptance threshold determines whether or not evidence is acceptable.

In MID, the values of the motivations are normalized so that the value of k is 1, and the importance index can be derived directly from the salient motivation.

Urgency, by contrast, is related to the situation itself, regardless of the importance that any individual attaches to it. In situations that demand immediate attention, the need to act is very strong. This is because urgency is related to action motivations which are high when salient. It is also related to importance, because urgent situations will increase the importance through heightened motivations. Importance and urgency each provide a complement to the other, one deriving its significance from the motivations of the ARA in a particular situation, the other from the situation itself. This notion of importance and urgency is similar to the idea of Maes [77] in which behaviour is determined both by motivations and goals on the one hand and by external observation on the other. It provides for the ability to respond based on both internal and external considerations.

| *importance* | *urgency* | $C_{action}$ | $C_e < C_{action}$ |
|---|---|---|---|
| high | high | mid-range | reject observations |
| high | low | high | reject observations and scenario |
| low | high | low | not possible |
| low | low | low | reject observations |

Table 6.3: Summary of rejection of evidence behaviour

### 6.5.3 Rejecting Evidence

In MID, the function, $f$, for combining the different parameters of uncertainty to obtain a value for *confidence* involves the multiplication of the four dimensions reliability, $r$, trustworthiness, $t$, accuracy, $a$ and credibility, $c$:

$$f(r,t,a,c) = r \times t \times a \times c$$

(Other functions are possible, but this is chosen for simplicity.) The values for confidence resulting from this function may be relatively low at an intuitive level. It should be stressed that the value itself is meaningful only so far as it determines whether or not evidence is acceptable, by comparison with the acceptance threshold and action point. Thus it would not be correct to say that a confidence value of 0.2 is low or that a value of 0.7 is high, but only that they are or are not acceptable. The meaning of the values is determined in relation to the motivations of the reasoning agent, and is high or low only in that respect.

Using the above function, MID determines whether the supplied evidence is acceptable or not. If MID finds that the evidence is not acceptable under the circumstances, then some element of it may be rejected. The rejection of evidence is strongly dependent upon importance and urgency. In situations with some urgency, only observations can be rejected, since the urgency of the situation demands that that particular situation be addressed. If urgency is not high, then with high importance, the scenario itself can be rejected if necessary. Rejection of the scenario is not usual, and is only allowed if there is no sufficiently reliable and trustworthy observer in the original scenario. MID's behaviour is summarized in Table 6.3. Note that when rejecting a scenario, the observations are also rejected. Note also that in the case of low importance and high urgency (higher than the limit for action thresholds), the confidence level cannot be lower than the action threshold value. Once MID finds that the evidence is acceptable it is used to reason about the current phenomenon.

### 6.5.4 An Example

Suppose we have a domain, $< \Delta, 0.9, 0.1, 0.7 >$ so that values of $C_{max}$ and $C_{min}$ are 0.9 and 0.1 respectively, and $\mathcal{U}_{lim}$ is 0.7. Now, given a scenario in which relatively accurate evidence can be observed, and which has no special features reducing its credibility, $< \mathcal{S}, 0.8, 0.9, 0.4 >$, (accuracy is 0.8, credibility is 0.9 and urgency is 0.4), and an observer who is relatively reliable but less trustworthy (values for reliability and trustworthiness of 0.9 and 0.8), we can calculate the confidence in the evidence. Using the simple multiplication function in MID, this is $(0.9 \times 0.8 \times 0.8 \times 0.9) = 0.52$.

The urgency of the situation (0.4) is below the limit for the domain (0.7), so the action point reduces to the acceptance threshold. Now, suppose that the salient motivation in MID has a moderately high strength value of 0.6 which determines the importance index. This lies between the confidence limits, so the acceptance threshold is simply the importance index which is 0.6. At this point, we know that the observations are not acceptable since the confidence (0.52) is below the acceptance threshold.

New observations are needed, and we can specify a requirement on the acceptable level of uncertainty that is introduced because of the observer. We know that accuracy and credibility are 0.8 and 0.9, and the acceptance threshold is 0.6, so we can require a combined value of not less than $0.6/(0.8 \times 0.9) = 0.83$ for the uncertainty due to the observer. Subsequent evidence is provided by a completely trusted observer with values of 0.9 for reliability and 1.0 for trustworthiness, giving a value of 0.9 for observer uncertainty, and 0.65 for combined confidence. The observations are therefore accepted. If it was not possible to exceed the threshold, then the scenario would have to be rejected and replaced with one whose values for accuracy and credibility would allow acceptance.

Note that the numbers themselves have no inherent meaning. They are significant only in determining whether evidence is acceptable, and whether action can be taken.

## 6.6 Discussion

### 6.6.1 Related Work

Most work on scientific discovery has considered the evaluation stage only implicitly. When it has been recognised, attention has been focussed merely on the simple comparison of observations with predictions in order to determine whether a failure has occurred

by simple matching. We will briefly consider how the systems described in Chapter 2 address evaluation. Rajamoney's COAST has no explicit evaluation component at all, and ignores the issues addressed here. The BACON systems by Langley et al. incorporate a minimal recognition of the importance of evaluation by using a simple fixed percentage error tolerance on the numerical data that is used. Also, BACON's architecture constrains the nature of the hypotheses that can be generated, so that no comparison of predictions and observations need be made, because only consistent hypotheses are considered. In Klahr and Dunbar's SDDS one of the three components of the top level architecture is devoted to evaluating evidence. They consider the rejection of the hypothesis, acceptance, or as a third possibility, the need to acquire more evidence and reassess the situation due to the current evidence being inconclusive. There is however, no elaboration as to what amounts to inconclusive evidence, and as mentioned earlier, no implementation. Reimann's HDD, while acknowledging the role of evaluation as one of his five steps, makes no distinction between *approximately* correct and *wrong* predictions, so that evaluation is again reduced to a simple comparison process.

The KEKADA system of Kulkarni and Simon, in attempting an accurate construction of historical discovery, allows a slightly more flexible evaluation. In the confidence measure that is attached to hypotheses, it includes the success and failure of the hypothesis as the number of experiments that *verify* or *falsify* it, and also the implied-success and implied-failure as the number of experiments which provide positive or negative indications but which are nevertheless inconclusive. The clear division of evidence into adequate and inadequate evidence in this way is attractive, but it is used in a rather different way to that proposed here. Measures of confidence are used to suggest promising hypotheses to explore or unpromising ones to discard.

Thagard's ECHO [118], a system which judges competing explanatory hypotheses on the basis of their coherence through a connectionist network, primarily addresses the problem of theory selection, and will be considered extensively in Chapter 8. In addition, however, it also relates data propositions to the hypotheses, and judges the coherence of the entire structure of data and hypothesis. Here, an evaluation of evidence is addressed. One of the principles on which the system is based, the principle of *data priority*, states that a proposition describing the results of observation has a degree of acceptability of its own. This seems to imply that evidence provided to the system is guaranteed to be true, and indeed, Thagard [118] explains that, "from past experience,

we know our observations are likely to be true." But, we also know that there are instances when our observations are not true, and we must address these. Although this is recognised through the deactivation of data units or propositions which cohere poorly with other propositions, it is an evaluation of evidence *after* the fact, and in relation to the hypothesis, not independently. As such, faulty data which coheres well with a hypothesis may be accepted. Callen [4], particularly with regard to legal reasoning, argues that ECHO's handling of evidence is thus inadequate, since it neither addresses seriously the possibility of the rejection of evidence, nor the possibility of gathering further evidence, a point made also by O'Rorke [83].

More recently, Zytkow's FAHRENHEIT system for integrated numerical discovery has been reorganised to include measuring instruments which can be used to investigate repeatability of results and measurement of error [135, 139]. By focussing initially on error determination through repeated base data, numerical error can be found and used to provide an error tolerance for the main experiments. The work is limited to numerical discovery, however, and the kind of error considered is limited to only one of the parameters of uncertainty discussed earlier, *accuracy*. Nevertheless, it provides an acknowledgement of the significance of error in machine discovery, and brings closer the possibility of applying discovery systems to real-world discovery problems.

In using testimonial evidence, Thost [127] has undertaken work on the combination of multiple, diverging evidence of testimony from different sources. This amounts to the attempted extraction of information from contradictory opinions by maintaining detailed knowledge of the factors determining the credibility of sources. The knowledge takes the form of Information Source Models (ISM) and includes details of competence, power (influence) and goals which are used to determine a source's credibility. Competence is similar to reliability discussed above, while goals encode a strength that approaches the notion of trustworthiness. Different modes provide different ways of combining the components of the ISM. Primarily, the ISM provides a way of representing a system's knowledge of its social context. Currently, MID does not maintain such knowledge, and relies on the source information to be provided with observations. Information Source Models could prove to be very useful if it was extended to do so, however.

Finally, it is worth noting that related research is being undertaken in modelling cooperation for autonomous agents. Marsh [78], for example, has developed a model for trust and reliance that is not dissimilar to the model described here for evidence

evaluation. Although the intention and application area are both different, the underlying concerns are with reasoning and acting effectively in an uncertain world.

### 6.6.2 Conclusions

More and more, as the difficulties of reasoning in the real world come to be appreciated, attention is turning to the evaluation of evidence in order to cope with noise and uncertainty. Reasoning systems, if they are to be considered effective, must be able not only to reason on the basis of supplied evidence, but also to reject evidence supplied to them if that evidence is inadequate. This is an important requirement, but one which has been neglected in the past. Moreover, the evaluation of evidence must consider, in addition to the evidence itself, an evaluation of the supplier or source of that evidence.

This chapter has described a model for evaluation which uses four parameters of uncertainty: accuracy, credibility, reliability and trustworthiness. The question of when evidence is acceptable has also been addressed, not in the form of a static, fixed rating of evidence, but by considering the evidence in the light of the motivations of the reasoning agent, and defining a variable measure for acceptability. The distinction between the model itself and its use in the MID system is deliberate, allowing other interactions with different implementations. Although the question of how importance may be defined is open, the role of importance is fixed. The decision to use a single salient motivation to determine importance rather than combining the complete set of motivations was taken for the sake of simplicity. Enhancements of this work may subsequently investigate other such possibilities, but the model holds regardless. Further work might extend the model by specifying the way in which the parameters of uncertainty can be updated based on the results of the reasoning or action that arises out of the use of accepted evidence.

The model provides a means of addressing the problem of evidence evaluation, which applies to reasoning in both centralized and distributed domains, and in both scientific and everyday contexts. It also relates to the problems associated with distributed sources of information which are increasingly being used in a variety of areas, and also to the modelling of autonomous agents in multi-agent environments. This is already the subject of investigation [78], and it seems that there is much benefit to be gained in attempting to combine the various approaches. The important point is that all evidence is susceptible to error and uncertainty. It must be evaluated in relation to the need for which it is obtained, and in such a way that allows a complete rejection of that evidence if necessary.

# Chapter 7

# Theory Revision

> ...the pencil's graphite is also the ephemeral medium of thinkers, planners, drafters, architects, and engineers, the medium to be erased, revised, smudged, obliterated, lost — or inked over.
>
> — Henry Petroski, *The Pencil: A History of Design and Circumstance*

## 7.1 Introduction

Revision is key to our model of inductive discovery. Indeed, much research into the problems of discovery comes under the heading of theory revision. Given an existing theory, new evidence will either be consistent with that theory or it will be anomalous. If it is consistent, then there is no cause for further reasoning since the theory is adequate. If the observations are anomalous, however, then the theory is refuted and must either be discarded or revised so that the anomaly is removed and the theory is once more consistent with observations. In the revision stage can be seen that part of the discovery cycle which is actually responsible for the construction of *new* theories. These new theories are limited in that they are derived solely from observations.

This chapter discusses exactly what revision entails, and examines the different kinds of revision that are possible. It begins by discussing very generally the problems of theory revision. Then it classifies revision into three groups and explains why each is necessary. This is followed with a more detailed investigation of revision operators, and how higher-order operators may be constructed from primitive ones. The next sections consider the kinds of knowledge that can be revised and the constraints that apply to the revision procedure. Finally, revision in MID is discussed, giving details of the revision strategy,

the order of revision, the revision operators that are used, and the algorithms for revision. The chapter ends by giving examples of the revision stage in MID, and by considering some related work. (Revision in MID is very similar to revision in Rajamoney's COAST system [91] because both use Qualitative Process Theory, though MID is more fully specified. The similarities and differences are explored towards the end of the chapter.) Further constraints on revision through selection are discussed in the following chapter.

## 7.2 The Revision Problem

Ginsberg [29] formulated the problem of theory revision as follows:

> A theory revision problem exists for a theory $\mathcal{T}$ when $\mathcal{T}$ is known to yield incorrect results for given cases in its intended domain of application. The goal of theory revision is to find a revision $\mathcal{T}'$ of $\mathcal{T}$ which handles all known cases correctly, makes use of the theoretical terms in $\mathcal{T}$, and may, with a reasonable degree of confidence, be expected to handle future cases correctly. In contrast to pure inductive learning from experience, theory revision is not only guided by the information implicit in $\mathcal{T}$, but also attempts to *preserve* the language and, as much as possible, the structure of $\mathcal{T}$.

He explains how this applies to theory revision in the sciences [30], the primary domain used as a basis for the model developed here. The given cases discussed above are merely associated experiments and observations. The cases that a theory handles incorrectly are thus instances of observations which cannot be accounted for by the theory. The new theory, $\mathcal{T}'$, in handling all known cases correctly, must be consistent with all prior observations, so that the two theories will be in agreement over that part of the domain for which the original theory, $\mathcal{T}$, is known to be accurate. Revising $\mathcal{T}$ to produce $\mathcal{T}'$ indicates a mapping from the old theory to the new theory such that the old theory is subsumed by the new one. This can be considered as the *reduction* of one theory to another. Making use of the theoretical terms of the old theory is a constraint that Ginsberg includes in order to avoid certain solutions such as the enumeration of cases or observations. Yet he acknowledges that this restriction must be relaxed at times so that theoretical terms can be dropped, and so that entirely new theories can be adopted.

In considering theory revision, Ginsberg seems to be addressing a number of issues at once without drawing important distinctions between the kinds of things to be done. There are a number of points that need to be made.

- Revision is the only stage of the model which can modify or construct knowledge. Any change to what is known can be considered a revision, even if it is not based upon previous knowledge. (This is so even if it is a Kuhnian revolution [51]. Ginsberg points out that the whole edifice of knowledge is never thrown out, even in these instances.) Thus new theories from external sources may be introduced at this point, and must be.

- Revision is necessary *only* when the theory is demonstrated to be inadequate in some way through failures to predict or explain observations.

- Revision must be guided by certain *implicit* constraints, but these should permit only those revisions that are warranted by the observations. Revisions which cannot be justified by appeal to observations alone should not be considered.

- Revision must also be guided by certain *explicit* constraints, but these are considered separately from the actual mechanisms of revision in the selection stage (see Chapter 8).

- Revision requires a finite but complete set of revision operators which transform an old theory into a new one through the application of these operators to remove anomalies.

## 7.3 Kinds of Revision and Why They are Necessary

The task of revising a domain theory is often regarded as a relatively straightforward operation. Given a domain theory (or any other body of knowledge), $\Delta$, represented in some language, $L$, there are only limited kinds of revision that can be performed. Following Levi [70], we note three kinds of revision to a body of knowledge that can be distinguished. These are shown in Table 7.1. (In fact, Levi actually considers a fourth kind of revision, residual shift, but this has received little attention in the literature, and Levi himself says nothing about it other than that it can be reduced to other kinds of revision, in which case we can ignore it.) Levi's concern in proposing this classification is in asking how it is that a revision to a body of knowledge that is considered consistent and correct can be justified. In considering this, an implicit constraint on what revision should allow is introduced.

| |
|---|
| **Expansion:** A shift is made from $\Delta_1$ to $\Delta_2$ containing $\Delta_1$ obtained by adding a sentence $e$ (or a set of sentences) to $\Delta_1$ and forming the deductive closure. |
| **Contraction** A shift is made from $\Delta_1$ to $\Delta_2$ where $\Delta_1$ is an expansion of $\Delta_2$. |
| **Replacement:** A shift is made from a consistent $\Delta_1$ containing $e$ to a consistent $\Delta_2$ containing $\neg e$. |

Table 7.1: Three kinds of revision.

### 7.3.1 Expansion

Expansion occurs when a reasoner adds information in the form of observations, laws or theories to a body of knowledge. The reason for doing this is simply that of adding new information itself. Demands for information will vary in degree and nature depending on the kind of reasoning undertaken, whether it is for the resolution of a practical and immediate decision or action problem, or is relevant to scientific reasoning aimed at explaining or predicting a particular phenomenon. The problem with expansion is that it introduces new information which may not be true into the body of knowledge which is considered correct and consistent at least so far as it is used for practical purposes[1]. Nevertheless, expansion is justified if any progress is to be made in the pursuit of satisfying goals which demand more or increasing knowledge, and if a paralysis of action is to be avoided.

### 7.3.2 Contraction

Contraction occurs when a reasoner removes information in some form from that body of knowledge. It is the inverse of expansion. The problem with this is that it deliberately decreases knowledge and, if there is a concern to obtain knowledge, then this is counter-productive. The justification for contraction can be seen in considering the nature of the real world when inconsistencies arise through errors in observation and experimentation, for example, despite strong efforts against this (see Chapter 6). Observations will thus sometimes contradict existing theories and may require the contraction of knowledge.

[1]This is a moot point. Philosophers such as Popper would disagree here, but effectively, acting on such knowledge can be regarded as an implicit acceptance of that knowledge (see earlier discussion in Chapter 6).

Contraction must aim to minimize the loss of information subject to the constraint that the inconsistency which demanded the contraction be removed.

### 7.3.3 Replacement

Replacement sometimes refers to the activity that Kuhn called a *scientific revolution* [51]. In this case, a theory (or part of a theory) is replaced by another which contradicts it. This is easily reduced to the first two kinds of revision by considering it simply as a contraction followed by an expansion. First, $T_1$ is removed from the body of knowledge, contracting, and then subsequent investigation leads to the addition of $T_2$, thus expanding the knowledge. The net result of the two revisions is replacement.

#### The role of observations

The classification of the different kinds of revision together with justifications reveals that observations are vital in constraining revision, both in the kind of revision to be used, and in the elements of knowledge that ought to be revised. If there is no inconsistency or incompleteness, then there is no need to revise. If there is a need to revise, then the revision must be constrained by the observations. This is not always the case, and it may certainly be legitimate to make revisions at least tentatively (this might be termed proposing hypotheses) based on factors other than observations alone, such as the use of analogy, for example (eg. [19]), but again this is outside the ability and scope of our concern here with reasoning in inductive discovery. The constraint of revising knowledge only as demanded by observations has been referred to as the *scenario constraint* [91] elsewhere and will be discussed further later.

## 7.4 Revision Operators

Revision operators are operators that can be applied to a body of knowledge (in this case a domain theory) to produce revised theories. If it is accepted that there are only two most basic kinds of revision and that all other kinds of revision can be reduced to some combination of them, then only two revision operators, corresponding to the two revision strategies of expansion and contraction, are needed. With little or no prior knowledge, these may be considered sufficient to achieve the appropriate results. It is unlikely that no applicable background knowledge will be available, however. Moreover, if none is

immediately available, then it is likely that some useful knowledge will be forthcoming over time through interactions in the relevant domain. This is important, for a lack of prior knowledge provides no useful constraints on the space of revisions that might be generated, and leads to *blind* explorations of the theory space. (Russell [102] discusses the different kinds of inductive learning possible, and the implications of reasoning with and without prior knowledge.) The benefit of prior knowledge is that more constraints on theory generation are available, and more sophisticated revision operators derived by combining the basic operators with this knowledge can be used.

### 7.4.1 Primitive Operators

At the most basic level, theories are either expanded or contracted. If they are expanded, then components are added. If they are contracted, components are deleted. We can therefore define two revision operators which revise a domain theory $\Delta$ to produce $\Delta'$:

**Addition** Some component $x$ is added to the domain theory.

$$\Delta' = \Delta \cup \{x\}$$

**Deletion** Some component $x$ is deleted from the theory.

$$\Delta' = \Delta \setminus \{x\}$$

Thus expansions and contractions are achieved by applying the addition and deletion operators to a component and things that follow from or imply that component respectively. (Note that the set notation is used here in a liberal way. A theory is more than just a collection of components, but the notation expresses the concepts of expansion and contraction through addition and deletion clearly and concisely. With stronger knowledge representation formalisms, these operators must be developed further.)

All possible revisions can thus be generated by using combinations of these operators alone. However, with appropriate background knowledge, we can go on to develop more sophisticated and potentially more powerful operators.

### 7.4.2 Higher-Order Operators

If existing knowledge can be brought to bear in choosing what to revise and how to revise it, then not only can blind revision be avoided in that the use of prior knowledge advances revision beyond mere syntactic manipulation, but a more complete and

coherent model of theory development arises that includes reference to a core of background knowledge (auxiliary assumptions, hypotheses and so on,) as well as the more immediately concerning object knowledge.

Typically, such background knowledge will be meta knowledge which relates different levels of description of objects in hierarchical structures and the like. Such knowledge allows the introduction of more revision operators, derived from the primitive operators, but which use this knowledge in proposing revisions which would not necessarily be possible from immediate empirical data alone. In particular, revision operators for specialization and generalization are highly desirable, commonly used, and facilitated by the use of background knowledge.

First though, it is important to consider one other revision operator which relies on background knowledge but less explicitly. Logical negation can be implemented using addition and deletion with background knowledge, but it is in some sense primitive, the background knowledge relevant here being that of logic. Negation indicates not merely a replacement, but the strong connection of opposition to what went before. The combination of addition and deletion according to observations does not adequately express the semantically significant change that would be made in such circumstances.

**Negation** Some component $x$ in the theory $\Delta$ is negated.

$$\Delta' = (\Delta \setminus \{x\}) \cup \{\neg x\}$$

**Specialization** Some component $x$ in the theory $\Delta$ is specialized.

$$\Delta' = (\Delta \setminus \{x\}) \cup \{S(x)\}$$

**Generalization** Some component $x$ in the theory $\Delta$ is generalized.

$$\Delta' = (\Delta \setminus \{x\}) \cup \{G(x)\}$$

More higher order operators are still possible through combining lower level operators in various ways, and through introducing more and more varied knowledge. Yet the operators considered here of addition, deletion, negation, specialization and generalization provide a limited, commonly used set of operators which are unlikely to require much combination and allow for simple revisions in most cases.

## 7.5 What to Revise

An important and relevant criticism of some of the more dogmatic proponents of falsificationism is that it is difficult to know exactly what has been falsified. Typically, a theory under investigation is subjected to experimentation so that inadequacies are revealed and the theory refuted and subsequently revised to be consistent with the new observations. This ignores the possibility that the error leading to the falsification is not in the theory itself (or that part of the theory under investigation), but in the background knowledge or auxiliary assumptions and hypotheses, the truth (or at least adequacy) of which is taken for granted. This is similar to the problem associated with observation. Observations cannot be relied upon without sufficient grounds for doing so, yet many accounts of the discovery process unquestioningly assume ideal observations and initial conditions. For the same reasons that ideal observations are denied, so is perfect background knowledge. In order to deal with this, background knowledge itself must be made amenable to revision. Yet if there is no distinction between background knowledge and object knowledge, then the the focus of the investigation is lost, distinctions between what is known and what is not are blurred, and everything loses certainty leading to a paralysis of action. This is unreasonable.

Background knowledge is knowledge that has been accepted as being adequate for using in reasoning processes with conviction, *unless* there are grounds for questioning it. There may be grounds for questioning it if no revisions can be found to accommodate anomalous observations, or perhaps if only *poor* revisions can be found. In this event, background knowledge may be revised. This might be considered a form of exception handling in that attention is turned away from the main theory only when no good solution (revision) is available.

This revision of background knowledge allows blame to be assigned outside the main theory causing erroneous revision to be avoided, and it also allows the further development of background knowledge as more information becomes available that is not part of the theory. There is however, no consensus over when it is acceptable for background knowledge to be revised, and it remains a difficult question.

## 7.6 Constraints on Revision

The number of possible revisions to a theory is huge. Without constraining the space of revisions, the problem of revision becomes intractable. In revising a theory, therefore, we must place many constraints on the revisions that can be generated, and these come in two forms which we can consider as implicit and explicit. *Implicit constraints* are constraints that cannot be varied in the course of the revision process. They are fixed and implicit in the sense that they are *hard* constraints which are used in order to make the space of revisions manageable and sensible in the light of observation. They are not addressed as part of the selection process, but are considered below. *Explicit constraints* are constraints which are explicitly specified and which are used to order the generation and subsequent selection of revisions. As such they will be considered in the next chapter on theory selection.

**Observation Constraint** The observation constraint is the most important, and most implicit constraint. It is designed to ensure that any revision that is generated must be sensible in terms of accounting for the observations that led to the failure of the domain theory. If there is no failure, then no revision is necessary. It is obvious that the observation constraint is necessary, but it is important to state explicitly as determining the kinds of revisions allowed.

**Scenario Constraint** The scenario constraint restricts the space of revisions that can be generated according to the context in which a failure is encountered, by requiring that only those revisions which are relevant to the current (failure) scenario are proposed [91]. This is a form of *simplicity* constraint which is considered further in Chapter 8. Of the many theories that can be generated, the scenario constraint prefers the simplest theories as those which involve only necessary modifications as demanded by the failure.

**Accuracy Constraint** Accuracy is problematic. (Here accuracy refers to the fit of the theory to the accepted evidence.) It is not always desirable to enforce the requirement of accuracy on the revision of theories since we may want to generate theories which are at most only approximately correct. Indeed, many of the theories which we commonly use are not accurate, but they provide an acceptable and effective means for dealing with our environment. Accuracy to some degree or

other, however, is ultimately necessary if we are to adhere both to the observation constraint and the scenario constraint above, for if there was no requirement of accuracy, then these constraints lose their strength and permit the generation of all manner of revisions.

**Selection Constraints** The selection constraints are the explicit constraints mentioned above. They serve to order the generation of revisions permitted by the above three constraints, and to select appropriate theories from those so generated. Selection constraints are considered in the next chapter.

## 7.7 Revision in MID

Let us now turn to how revision is implemented in MID. This section gives a brief review of the knowledge structures used in the system, and then moves on to specify the revision operators that may be used to modify them.

### 7.7.1 Review of Knowledge Structures

There are two main knowledge structures that MID is could revise: the domain theory that is the repository for explicit knowledge of the phenomena under investigation and the background knowledge that is used in tandem with the domain theory in order to draw inferences. The domain theory is represented in a variant of Forbus' Qualitative Process Theory (QPT) [22], while the background knowledge is represented in the form of implication rules.

#### Domain Theory

Figure 7.1 shows the contents of an example domain theory, together with an expanded process. In revising the domain theory, we can identify two main components: conditions and effects. Using this distinction allows strong parallels to be drawn between this system and others based on alternative representations. At a lower level, the conditions include the slots of *individuals*, *preconditions* and *quantity conditions*. The effects include the slots of *relations* and *influences*. Each of these slots can be revised as described below.

```
Process:  solution
Process:  evaporation
Process:  condensation
Process:  absorption
Process:  release

Process Name:           fluid-flow
Individuals:            contained-fluid ?source
                        contained-fluid ?destination
                        path ?path
Preconditions:          path-connection ?source ?destination ?path
                        flow-aligned ?path
QuantityConditions:     greater-than (a (pressure ?source))
                                     (a (pressure ?destination))
Relations:              Q + fluid-flow-rate (pressure ?source)
                        Q - fluid-flow-rate (pressure ?destination)
Influences:             I + (amount-of ?destination) (a (fluid-flow-rate))
                        I - (amount-of ?source) (a (fluid-flow-rate))

Process:  add-solute
```

Figure 7.1: An abbreviated Qualitative Process domain theory for MID

```
flow-aligned          →  aligned
fluid-flow-aligned    →  flow-aligned
heat-flow-aligned     →  flow-aligned
```

Figure 7.2: An example background knowledge rule base.

### Background Knowledge

Figure 7.2 shows a very restricted example background knowledge rule base. Each rule has an antecedent and consequent, and represents the implication of the consequent by the antecedent. Either part of the rule could be revised, but it should again be noted that revisions to the background knowledge should *only* be possible in exceptional circumstances.

## 7.7.2 Kinds of Anomaly

Evaluation involves the comparison of predictions generated from the domain theory with observations provided from outside the system. In the case of a successful prediction when an observation is expected, no further work is necessary. If, however, an anomaly results, then the domain theory must be revised accordingly. This requires that the cause of the failure be identified in order that it may be corrected. We note three possible

| | | |
|---|---|---|
| **Case 1:** | $O = \{o_1\},\ P = \{p_1\},\ o_1 = p_1$ | *No failure* |
| **Case 2:** | $O = \{o_1\},\ P = \{p_1, \ldots, p_n\},\ o_1 \notin P$ | *Anomalous Observation* |
| **Case 3:** | $O = \{o_1, \ldots, o_n\},\ P = \{p_1\},\ p_1 \notin O$ | *Anomalous Prediction* |

Table 7.2: Three possible results of evaluation.

cases: no failure, anomalous observation, and anomalous prediction. In the first case, the predictions are consistent with the observations and no failure results. The two kinds of anomaly are as follows:

**Anomalous Observation** An observation that was observed was not predicted. This is sometimes referred to as an *unexpected observation.*

**Anomalous Prediction** A prediction that was generated by the inference mechanism was not observed. It is sometimes referred to as a *failed prediction.*

All three cases are specified in Table 7.2, where $O$ is the set of observations, and $P$ is the set of predictions.

For a set of observations and predictions, we must address all of the anomalies. MID considers all of the anomalous observations first, and then the anomalous predictions. For each observation in turn, the domain theory, $\Delta$, is revised accordingly. So for example, say we have $O = \{o_1, o_2, o_3\}$ and $P = \{p_1, p_2, p_3\}$, and $o_1 = p_2, o_2 \notin P$, and $o_3 \notin P$, then revision is performed as follows:

1. $O = \{o_1, o_2, o_3\}, P = \{p_1, p_2, p_3\} \mid o_2 \notin P$ *Anomalous Observation* **(Case 2)**
   Revise the domain theory.

2. $O = \{o_1, o_2, o_3\}, P = \{p_1, p_2, p_3, p_4\} \mid o_2 = p_4, o_3 \notin P$ *Anomalous Observation* **(2)**
   Revise domain theory.

3. $O = \{o_1, o_2, o_3\}, P = \{p_1, p_2, p_3, p_4, p_5\} \mid o_3 = p_5, p_1 \notin O$ *Anomalous Prediction* **(3)**
   Revise domain theory.

4. $O = \{o_1, o_2, o_3\}, P = \{p_2, p_3, p_4, p_5\} \mid p_3 \notin O$ *Anomalous Prediction* **(3)**
   Revise domain theory.

5. $O = \{o_1, o_2, o_3\}, P = \{p_2, p_4, p_5\} \mid O = P$ *No failure* **(1)**
   No action.

However, the resolution of one anomaly may lead to the possibility of causing a new anomaly to be introduced or, alternatively, the resolution of a later anomaly may cause an earlier anomaly (already eliminated) to be re-introduced. If we regard each anomaly as a goal that must be achieved, then the revision procedure is analogous to the generation of plans. Just as the interactions between goals can cause problems in the planning process, so too can the anomalies in the revision process here. By constraining the revision process so that we generate only those revisions which can be consistent with all of the current observations, the consistency of the revised theories is assured. This is possible because there is no temporal ordering on anomalies, and we know exactly what the final revised theory must allow.

In fact, this requirement could be relaxed, because through further revision, anomalies that were introduced as part of the revision process could subsequently be eliminated. This would, however, lead to the possibility of infinite recursion with new anomalies being introduced at each stage of revision. The constraint also enforces a degree of *conservatism* (discussed in the next chapter) that prefers revisions with fewer individual modifications to those which require many modifications to be consistent.

### The Grouping Heuristic

The view of revision described here is somewhat simplified. In attempting to minimize the amount of effort devoted to revision, MID groups observations together and tries to perform revisions based on the group as a whole. If a single revision can account for a group of three anomalous observations, for example, as opposed to three separate revisions if they were addressed individually, then the saving is substantial. If it cannot account for them, then progressively smaller groups of observations are used until the base case of individual observations is used. This also implements a form of *conservatism* by which the simplest revisions are attempted first, and more complicated revisions are tried only if necessary. With increasingly large numbers of observations, this strategy offers considerable benefit.

MID groups observations in the order in which they are presented to it, and uses successively smaller groups. This excludes groups of certain combinations of observations, but it reduces the number of groups that must be tested, so that attempts to optimize performance are not compromised by exhausting each possiblility. More complete combinations are possible by specifying different groupings, the most complete being achieved

by using the power set of observations as input for revision. (Appendix A gives an example showing how MID generates revisions with and without the *grouping* heuristic.)

### 7.7.3 Revision Operators

The observation constraint requires that all revisions must be directed at eliminating the anomalies that can arise. In the case of anomalous prediction, this means that either an active process which caused the prediction to be generated must be made to be inactive by modifying its conditions, or the effects of an active process which caused the prediction must be modified so that the rogue predictions are not generated. In the case of anomalous observation, the converse is true: either an inactive process which would have generated the predictions must be made active through modification of its conditions, or an active process must be made to predict the observations through modification of its effects. An extra possibility is that a new process can be created which predicts the observations.

Below we specify the revision operators that are required in MID. Most of the operators function as expected (eg., addition and deletion), but where they do not they are illustrated with examples.

It is important to note that while all of these revision operators can be reduced to the different kinds of revision considered earlier, the relation between them may not be obvious. Because of the frame structure of the knowledge representation, *adding* a condition will have the effect of reducing the scope of a process, resulting in a *contraction*, and not an *expansion* as might be expected. Similarly for deleting a condition, the results may be counter intuitive.

#### Revise Individuals

Individuals specify the type of the variable concerned. They can be considered to be conditions because they provide constraints on the applicability of a process in that they require instantiation as participants, without which a process cannot be activated. However, they have a very different status from the other kinds of condition. The predicates which specify type values cannot be negated. Individuals are only revised if this is needed by revisions to other components of a process. They cannot be revised independently[2].

[2]Currently, the revision of individuals in MID is limited. In particular, specialization and generalization of individuals are not implemented because of the distinct *type* nature of the predicates. These

**Add-Individual** This adds an individual to a process. Individuals are added only if they are needed in other parts of the process through the addition of other kinds of conditions or effects.

**Delete-Individual** This deletes an individual from a process. Individuals are deleted only if they are not required in other parts of the process.

**Specialize-Individual** This uses the background knowledge to find an appropriate specialization of the individual predicate. Individuals can only be specialized if they are appropriate to the corresponding conditions and effects. It is used in the same way as **Specialize-Pcondition**, for which an example is given below.

**Generalize-Individual** This finds an appropriate generalization of the individual predicate. Again, the generalization must be appropriate in light of the corresponding conditions and effects. See below for an example of generalization.

### Revise Conditions

There are two kinds of condition slot: preconditions and quantity conditions, and we will consider these in turn.

### Preconditions

Preconditions specify qualitative conditions on the applicability of a process through predicates and variables. They allow more scope for revision than individuals because the predicates are not simply type declarations as they are for individuals. (We abbreviate to Pcondition.)

**Add-Pcondition** This adds a new precondition predicate together with its associated variables to the precondition slot.

**Delete-Pcondition** This deletes a precondition.

**Negate-Pcondition** This negates the predicate of a precondition. Negation involves adding a not, so that we might modify the first precondition of Figure 7.1, giving:

```
not-path-connection ?source ?destination ?path
flow-aligned ?path
```

predicates are held to be primitive. Specialization and generalization are possible, however, by moving relevant predicates to the preconditions slot. The extra operators are specified here for completeness.

**Specialize-Pcondition** This uses the background knowledge to find an appropriate specialization of the precondition predicate by traversing the rules in the background knowledge rule base. For example, we might specialize the second precondition of Figure 7.1, using the second rule in Figure 7.2:

```
path-connection ?source ?destination ?path
fluid-flow-aligned ?path
```

**Generalize-Pcondition** This uses the background knowledge to find an appropriate generalization. For example, we might generalize from Figure 7.1 using the first rule, giving:

```
path-connection ?source ?destination ?path
aligned ?path
```

### Quantity conditions

Quantity conditions specify quantitative conditions on the applicability of a process through `greater-than`, `less-than` and `equal-to` predicates. Quantity conditions cannot be generalized or specialized. (We abbreviate here to Qcondition.)

**Add-Qcondition** This adds a new quantity condition comprising a quantitative predicate and a pair of variables expressing a quantitative constraint to a process. Two kinds of quantity condition are possible: quantity conditions which specify maximum limit points on increasing quantities and minimum limit points on decreasing quantities; and quantity conditions which specify relationships between quantities which move in opposite directions. To Figure 7.1 we can add :

```
greater-than  (a (amount-of ?source))
              (a (minimum-amount-of-point ?source))
less-than     (a (amount-of ?destination))
              (a (maximum-amount-of-point ?destination))
```

in the first case, and in the second case we add:

```
less-than (a (amount-of ?destination)) (a (amount-of ?source))
```

**Delete-Qcondition** This deletes a quantity condition.

**Negate-Qcondition** This negates the predicate of a quantity condition in the expected way, by replacing `greater-than` with `less-than` *and* `equal-to`, and so on. Negating the quantity condition of Figure 7.1, for example, gives two possible revisions:

```
less-than (a (pressure ?source)) (a (pressure ?destination))
```
*and*
```
equal-to (a (pressure ?source)) (a (pressure ?destination))
```

### Revise Effects

There are two effects slots which can be modified, direct effects as influences, and indirect effects as relations.

### Relations

Relations specify indirect effects of a process. These are effects which specify relationships between quantities, but which are not influences[3].

**Add-Relation** This adds a new relation to a process.

**Delete-Relation** This deletes a relation from a process.

**Invert-Relation** This is analogous to negating the relation. It modifies it so that the predicted effect is the inverse (or opposite) of the original effect. See **Invert-Influence** for an example.

**Specialize-Relation** This narrows the scope of an influence by replacing the whole of a quantity by a part of it. See **Specialize-Influence** below.

**Generalize-Relation** This replaces a part of a quantity by the whole quantity.

### Influences

Influences specify effects which are direct results of the process. They specify relationships between quantities and the process itself.

**Add-Influence** This adds an influence to a process.

**Delete-Influence** This deletes an influence from a process.

**Invert-Influence** This inverts the influence so that the modified influence is the inverse of the original influence. If we were to invert the second influence above, we would get both source and destination increasing:

[3]Relation revision operators have not been implemented in the current version of MID.

```
I+ (amount-of ?destination) (a (fluid-flow-rate))
I+ (amount-of ?source) (a (fluid-flow-rate))
```

**Specialize-Influence** This narrows the scope of an influence by replacing the whole of a quantity by a part of it[4]. For example we could introduce the notion of solutes into the previous case, and could get:

```
I + (amount-of ?destination) (a (fluid-flow-rate))
I + (amount-of:solute-of ?source) (a (fluid-flow-rate))
```

**Generalize-Influence** This replaces a part of a quantity with the whole quantity. Generalization would simply reverse the specialization.

### Creating a New Process

A new process can be created which instantiates each slot according to the current situation. The individuals slot is filled with variables that are used elsewhere in the process, the precondition slot is filled with a predicate that characterizes the current situation, the influence slot is filled with the observations. All other slots are left empty.

### Assessment

Rajamoney [91] defines a taxonomy of operators for COAST that is similar to that described here. This departs from his in a number of ways, however, through all levels. At the topmost level, Rajamoney has a third possibility, namely inverse behaviour, but in MID this reduces to a combination of anomalous observation and anomalous prediction failures. At the lower levels, many operators are the same since the representation is largely the same, but they are developed further in MID. In particular, it is not clear how specialization and generalization take place, since no mention is made of background knowledge that might be used to provide the classificatory information necessary. Furthermore, individuals cannot be added or deleted from processes. COAST is discussed further in the section on related work at the end of this chapter.

Tables 7.3 and 7.4 specify which operators may be used in the case of anomalous predictions and anomalous observation failures respectively. Note that operators on

[4]Currently, MID does not hold knowledge of the part to whole structure of objects and quantities. The specialize and generalize operators for influences are consequently not implemented in the current version. They are included here for completeness of specification, however.

| Anomalous Prediction | |
|---|---|
| *revise conditions* | *revise effects* |
| Add-Pcondition | Delete-Influence |
| Negate-Pcondition | Invert-Influence |
| Specialize-Pcondition | Specialize-Influence |
| Add-Qcondition | Delete-Relation |
| Negate-Qcondition | Invert-Relation |
| | Specialize-Relation |

Table 7.3: Revision operators for anomalous prediction failures.

| Anomalous Observation | | |
|---|---|---|
| *revise conditions* | *revise effects* | *new process* |
| Delete-Pcondition | Add-Influence | New-Process |
| Negate-Pcondition | Invert-Influence | |
| Generalize-Pcondition | Generalize-Influence | |
| Delete-Qcondition | Add-Relation | |
| Negate-Qcondition | Invert-Relation | |
| | Generalize-Relation | |

Table 7.4: Revision operators for anomalous observation failures.

individuals are not included since they are used to facilitate the application of another operator on some other element of the process.

### Revising Background Knowledge

By making the background knowledge structures explicit and accessible, the rules in the background knowledge base can be revised. No prior knowledge can be used when revising the rule base. Moreover, because it is held with a greater confidence than the domain theory, the revisions that can be made to the rule base may be limited in allowing only one revision at a time. This enforces a *conservatism* constraint reflecting the status of the strength and confidence of the BKRB. Consequently, we consider only two operators.

**Add-Rule** This adds a rule to the background knowledge base in attempting to allow a generalization of a precondition to be made.

**Delete-Rule** This deletes a rule from the background knowledge in attempting to prevent a generalization from being made.

Thus rules would be added in the case of an anomalous observation failure, and rules would be deleted in the case of an anomalous prediction.

Revising background knowledge (theory) is, however, as mentioned earlier, a very radical step to take. If background knowledge is revised, then the basis on which the development of the theory takes place is brought into question. It might be argued that the revision of background knowledge invalidates the existing theory entirely, so that new background knowledge may require a new theory to be constructed from the beginning. (This could be seen as a revolution in Kuhn's terms.) Moreover, in the MID system, the background knowledge rule base is used in a definitional way, grounding high level predicates in lower level ones. Changes to such knowledge are thus much harder to justify because the implications of these changes can be very significant for the existing domain theory. At an implementation level, modifying a rule-base is a difficult task. If the rule-base is to be useful, then any change must ensure that consistency is preserved throughout its deductive closure. In general, with large background knowledge rule bases, ensuring consistency may demand excessive resources, and be intractable. Because of these difficulties, the BKRB revision operators are not currently implemented.

Circumstances when revision of the background knowledge may be an acceptable solution are discussed further in the next chapter.

### 7.7.4 Algorithms for Revision

This section presents details of the algorithms of the revision process. Since all but one (New-Process) of the different kinds of revision are similar, an abstract algorithm is given first, followed by details of each of the separate cases of revision. In order to specify the separate cases, we need to introduce some more notation. The notation used in algorithm specifications that follow is summarized in Table 7.5.

The basic algorithm for revision has three main parts. These are enumerated in Table 7.6. For each class of revision the algorithm is elaborated subsequently.

#### Anomalous Prediction

There are two possibilities for anomalous prediction: condition revision and effect revision. Table 7.7 specifies the instantiation of the abstract algorithm for both of these.

In the case of condition revision, a process that caused the failed predictions to be generated is revised so that it is no longer active in the current scenario, thus retracting

| |
|---|
| $\Delta$ is the domain theory — the set of all processes in the theory. |
| $\Delta'$ is the revised domain theory. |
| $X$ is the set of anomalies currently being addressed. |
| $H$ is the set of processes (hypotheses) applicable to the revision. |
| $H'$ is the set of revised processes. |
| $P_h$ is the set of predictions that can be generated from the process $h$. |
| $\mathcal{R}$ is the set of revision operators applicable. |
| $Ap$ is the set of anomalous predictions. |
| $Ao$ is the set of anomalous observations. |
| $Ap'$ and $Ao'$ are the sets of anomalies updated after revision. |

Table 7.5: Notation used in this section.

---

**Basic Revision Algorithm**

1. Retrieve all of the processes in the domain theory to which the current kind of revision can apply. This entails specifying $H$.

2. For each such process, generate all possible revisions that are warranted by the observations and the scenario description. Here, $H'$, the set of revised processes is determined by specifying the set of operators, $\mathcal{R}$ used to produce revised processes, $\mathcal{R}h$. $H'$ is defined as follows:

$$H' = \{h' \mid \exists h \in H \cdot \exists \rho \in \mathcal{R} \cdot h' \in \rho h\}$$

3. Update accordingly the lists of anomalies still to be resolved.

4. Update the domain theory by replacing modified processes so that

$$\Delta' = (\Delta \setminus \{h\}) \cup \rho h$$

---

Table 7.6: The specification of the abstract revision algorithm.

**Condition Revision**

1. $H = \{h \mid \text{active}(h) \wedge \forall x \in X \cdot x \in P_h \wedge P_h \subseteq Ap\}$
2. $\mathcal{R}$ = **{Add-Pcondition, Negate-Pcondition, Specialize-Pcondition, Add-Qcondition, Negate-Qcondition}**
3. $Ap' = Ap \setminus P_h$

**Effect Revision**

1. $H = \{h \mid \text{active}(h) \wedge \forall x \in X \cdot x \in P_h\}$
2. $\mathcal{R}$ = **{Delete-Effect, Invert-Effect, Specialize-Effect}**
3. $Ap' = Ap \setminus X$

Table 7.7: Revision algorithm for anomalous prediction failures.

the predictions. Condition revision can apply to all processes which cause the current anomaly, and which do not cause correct predictions. The operators are applied to generate revisions $\mathcal{R}h$ so that only *one* condition is unsatisfied. (This could be more than one, but conservatism and simplicity counsel revision only as much as necessary.)

With effect revision, the effects of a process that caused the failed prediction are revised so that the failure is avoided. Effect revision applies to all active processes which cause the current anomaly.

### Anomalous Observation

There are three possiblities for anomalous observation: condition revision, effect revision, and new process. Table 7.8 specifies the instantiation of the abstract algorithm for the condition and effect revisions.

In the case of condition revision, the conditions of a process that prevented it from being active are revised so that it can predict the observations. Condition revision applies to processes which could prevent the anomaly if they were active. The operators are combined appropriately so that *all* of the conditions are satisfied.

In effect revision, the effects of active processes are revised so that they cause the anomalous observation. Effect revision applies to active processes with variable bindings that include the anomalous observation quantities.

**Condition Revision**

1. $H = \{h \mid \text{inactive}(h) \wedge \forall x \in X \cdot x \in P_h \wedge (P_h \subseteq Ao)\}$
2. $\mathcal{R}$ = **{Delete-Pcondition, Delete-Qcondition, Negate-Pcondition, Negate-Qcondition, Generalize-Pcondition}**
3. $Ao' = Ao \setminus P_h$

**Effect Revision**

1. $H = \{h \mid \text{active}(h) \wedge \forall x \in X \cdot x \notin P_h \wedge \text{vars}(x) \subseteq \text{vars}(h)\}$
2. $\mathcal{R}$ = **{Add-Effect, Generalize-Effect, Invert-Effect}**
3. $Ao' = Ao \setminus X$

Table 7.8: Effect and Condition revision algorithms for anomalous observation failures.

The new-process revision is somewhat different. Here, a new process is created that accounts for the anomalous observations. This is specified in Table 7.9.

**New Process**

1. One new process only, $h$, is created that accounts for *all* anomalous observations.
2. All remaining anomalous observations are removed, so that $Ao' = \emptyset$
3. The new process is added to the domain theory: $\Delta' = \Delta \cup h$

Table 7.9: The new-process algorithm for anomalous observation failures.

### 7.7.5 A Simple Example

To illustrate the revision procedure, consider the domain theory specified by Figure 7.3. The theory contains knowledge about only one *process*, heat-flow, and there is no background knowledge. If the conditions are satisfied, then the theory predicts that the temperature of the destination object will increase. The theory is erroneous in that it does not know about the other effect of heat flow, that the temperature of the source will decrease. Now, say MID is provided with the scenario description of Figure 7.4, then the heat-flow process will be active since all of the conditions are satisfied. Accordingly,

```
Process Name:          heat-flow
Individuals:           object ?source
                       object ?destination
                       heat-path ?path
Preconditions:         heat-connection ?source ?destination ?path
                       heat-flow-aligned ?path
QuantityConditions:    greater-than (a (temperature ?source))
                                    (a (temperature ?destination))
Relations:             Q+ heat-flow-rate (temperature ?source)
                       Q- heat-flow-rate (temperature ?destination)
Influences:            I+ (heat ?destination) (a (heat-flow-rate))
```

Figure 7.3: An erroneous domain theory concerning heat flow

```
Scenario Name:   heat-flow-works-scenario
Individuals:     std-object1
                 std-object2
                 std-path1
Facts:           object std-object1
                 object std-object2
                 heat-path std-path1
                 heat-connection std-object1 std-object2 std-path1
                 heat-flow-aligned std-path1
                 greater-than (a (temperature std-object1))
                              (a (temperature std-object2))
```

Figure 7.4: A scenario description in which heat flow occurs

MID makes the appropriate prediction shown in Figure 7.5.

### Anomalous Observations

If MID is presented with the observations:

```
> increase (heat std-object2)
> decrease (heat std-object1)
```

then there is an anomalous observation because MID did not predict the latter. MID therefore attempts to revise its domain theory. There are two possibilities. First, the effects of active processes can be modified so that the unexpected observation is included.

```
Predictions:  increase (heat std-object2)

I+ (heat std-object2) (a (heat-flow-rate))
 Active:  heat-flow std-object1 std-object2 std-path1
          heat-connection std-object1 std-object2 std-path1
          heat-flow-aligned std-path1
          greater-than (a (temperature std-object1)) (a (temperature std-object2))
```

Figure 7.5: The predictions generated by MID

```
PROCESS:        heat-flow
Variables:      ?source
                ?destination
                ?path
Individuals:    object ?source
                object ?destination
                heat-path ?path
Pconditions:    heat-connection ?source ?destination ?path
                heat-flow-aligned ?path
Qconditions:    greater-than (a (temperature ?source))
                             (a (temperature ?destination))
Relations:      Q+ (heat-flow-rate) (temperature ?source)
                Q- (heat-flow-rate) (temperature ?destination)
Influences:     I+ (heat ?destination) (a (heat-flow-rate))
                I- (heat ?source) (a (heat-flow-rate))
Revision Log:   1 add influences
```

```
No change:      heat-flow

PROCESS:        process2
Variables:      ?var-4
Individuals:
Pconditions:    precondition-heat-flow-works-scenario ?var-4
Qconditions:
Relations:
Influences:     I- (heat ?var-4) (a (process2-rate))
RevisionLog:    1 new-process
```

Figure 7.6: The revisions generated by MID for the anomalous observation example

Second, a new process that accounts for the unexpected observation can be added. These revisions are shown in Figure 7.6.

### Anomalous Predictions

Alternatively, if we have the same theory, the same scenario, and the same predictions, but cannot observe any changes in the world, then we get an anomalous prediction failure. In this case, we can either revise the process in the theory by modifying its conditions so that it is not active, or we modify its effects so that it does not cause unobserved predictions. Figure 7.7 shows the revisions that MID generates here.

The first revised theory is given completely, but due to space constraints, subsequent revisions are specified only by the slots that have changed from the original. Although the second example here is somewhat contrived, it illustrates the nature of the revision procedure, and the kind of revisions that MID generates in the appropriate circumstances. Even with such a small domain theory comprising only a single process definition, the number of revisions that can be generated is sizeable, and some means of choosing be-

| | |
|---|---|
| PROCESS: | heat-flow |
| Variables: | ?source |
| | ?destination |
| | ?path |
| Individuals: | object ?source |
| | object ?destination |
| | heat-path ?path |
| Pconditions: | heat-connection ?source ?destination ?path |
| | heat-flow-aligned ?path |
| | precondition-3 ?source ?destination ?path |
| Qconditions: | greater-than (a (temperature ?source)) |
| | (a (temperature ?destination)) |
| Relations: | Q+ (heat-flow-rate) (temperature ?source) |
| | Q- (heat-flow-rate) (temperature ?destination) |
| Influences: | I+ (heat ?destination) (a (heat-flow-rate)) |
| Revision Log: | 1 add preconditions |
| PROCESS: | heat-flow |
| Pconditions: | not-heat-connection ?source ?destination ?path |
| | heat-flow-aligned ?path |
| Revision Log: | 1 negate preconditions |
| PROCESS: | heat-flow |
| Pconditions: | heat-connection ?source ?destination ?path |
| | not-heat-flow-aligned ?path |
| Revision Log: | 1 negate preconditions |
| PROCESS: | heat-flow |
| Qconditions: | greater-than (a (temperature ?source)) |
| | (a (temperature ?destination)) |
| | greater-than (a ((heat ?destination))) |
| | (a (minimum-heat-point ?destination)) |
| Revision Log: | 1 add quantityconditions |
| PROCESS: | heat-flow |
| Qconditions: | equal-to (a (temperature ?source)) |
| | (a (temperature ?destination)) |
| Revision Log: | 1 negate quantityconditions |
| PROCESS: | heat-flow |
| Qconditions: | less-than (a (temperature ?source)) |
| | (a (temperature ?destination)) |
| Revision Log: | 1 negate quantityconditions |
| PROCESS: | heat-flow |
| Influences: | |
| Revision Log: | 1 delete influences |

Figure 7.7: Revisions generated by MID for the anomalous prediction example

tween the different revisions is required. With increasing numbers of processes in the domain theory, the number of potential revisions will increase proportionately. It is unlikely that empirical evidence will always be available to discriminate between competing theories, and we must therefore look to other criteria for selecting one appropriately.

## 7.8 Discussion

### 7.8.1 Related Work

Most of the systems introduced in Chapter 2 have only limited revision abilities. This is particularly true for BACON, STERN and HDD which concentrate on trend detection in the discovery of numerical laws. HDD does have a mechanism for attaching conditions to hypotheses in the event of failure, but this is very limited. There are a couple of systems, however, which do address revision in a similar way.

Rajamoney's work on COAST in particular [91] is very closely related to the work here since it also uses Qualitative Process Theory to represent its domain theory. He uses a scheme of abstraction in revision by which groups of proposed revisions are abstracted at a high level so that they may be subject to experimentation collectively. This allows the refutation of many revised theories before they are completely generated, and relies heavily on the experimentation component of the system. Abstract revisions which cannot be differentiated on the basis of experiments are then refined to concrete theories. The set of revision operators that is used by COAST, however, is not complete. It does not allow individuals to be added or deleted from processes, and it does not allow individuals to be modified in any way either.

Furthermore, the revision strategy adopted by COAST is very restricted and this also constrains the potential revisions so that only a subset can be generated. COAST groups observations together and considers only revisions which can account for the whole group at once, ignoring revisions that involve multiple individual revision operator applications. By contrast, MID groups observations together in order to simplify the revision procedure, and to conserve resources. If no revision can be generated, MID progressively considers more complicated revisions all the way through to those using observations individually. Thus we have considered a complete set of revision operators, constrained only by what is syntactically and semantically acceptable.

Finally, it should be noted that Rajamoney does not allow the possibility of modifi-

cation of background knowledge.

Karp's HYPGENE program [46], which reasons in the field of molecular biology, also uses a qualitative representation to model the domain theory, and provides a similar set of operators to those described here. He also mentions the possibility of including operators for modifying the class knowledge base of the system (similar to the background knowledge base here), but provides no details of how this might be implemented.

An analogous problem to theory revision is that of knowledge base refinement which involves the modification of a knowledge base as opposed to a (scientific) domain theory (see, for example, [28, 31]). Ginsberg [30] points out that much of this work on theory revision applies equally to knowledge base refinement. Both involve the modification of a repository of knowledge in response to failures or inadequacies of some kind. We will not attempt a detailed analysis of the similarities or differences here, however.

### 7.8.2 Conclusions

Theory revision is a systematic process. Without the use of selection criteria to constrain the space of revisions (which we consider in the next chapter), it is reduced to the syntactic manipulation of domain theories. This in itself requires that a complete set of revision operators exists, something which has been lacking in some previous systems. There are, however, constraints. Revised theories must be consistent with the observations and scenario that caused the theory failure to arise, and the revision operators must therefore be appropriately applied. Furthermore, the use of background knowledge in some form in order to facilitate the revision of theories, provides other requirements. The background knowledge itself should be capable of being revised, at least in principle, for it is not inviolable, merely accepted with a greater certainty. Thus it should be more difficult to revise background knowledge, but not necessarily impossible.

MID provides a facility that allows effective and efficient theory revision. It is similar to the revision procedures in other systems that use similar knowledge representations, but offers significant advances. By grouping observation, revisions are generated in a progressively more complicated way, so that the simplest revisions are considered first, and complicated revisions later. While other systems consider only simple revisions, MID retains this advantage in terms of resources, but makes the procedure complete by considering more complicated revisions as and when necessary.

# Chapter 8

# Theory Selection

> ...there is trouble in store for anyone who surrenders to the temptation of mistaking an elegant hypothesis for a certainty ...
>
> — Primo Levi, *The Periodic Table*, Chromium

## 8.1 Introduction

Theory selection is the problem of choosing a 'best' theory from a large number of theories or potential theories. There are infinitely many theories that can be constructed to explain a particular observed phenomenon, and some means is required for discriminating between them. This entails the use of various criteria by which the merit of theories can be judged. Through these criteria, a *bias* (in a broad sense) on theories can be imposed, resulting in useful and effective theories which satisfy needs for generalization.

Biases may be imposed in many ways. The bias imposed by the representation language used to express theories has already been discussed, and provides a significant restriction on the space of potential theories. Chapter 7 introduced a number of other implicit biases which cannot be altered in the MID system. These are imposed to restrict the space of allowable theories to those which address the requirements of the system of accounting for observations and allowing predictions to be made. Many theories may still be generated, however, and it is not necessarily possible, solely on the basis of experiment and observation, on the basis of empirical evidence, that a unique theory choice may be made. MID uses heuristic search through the space of revisions to impose a bias that informs this choice. These heuristics are criteria for the virtue and acceptance of theories. Many such criteria have been proposed, mostly in the field of the philosophy of science,

but some have been extended to a computational context. In abduction, which can be considered to be *inference to the best explanation* [36], can be seen similar concerns. Just as the theory is selected according to certain criteria, so is the explanation derived from that theory. We note the relevance of work on abduction, but do not explicitly consider it further. Much of this chapter, however, is equally applicable and relevant to abduction.

In addressing the problems of selection, the following points must be considered:

- Theory selection is not an independent process. It is intimately connected to the process of generation or revision of theories, and an adequate account of selection must consider not only the selection of theories that are supplied by an outside source, but also using selection to guide and constrain the generation or revision process that provides the theories in the first place.

- Selection is dependent on the motivations of the reasoner. What might be appropriate for one reasoner may not be so for another with different motivations and priorities. Selection must address the issues of variation of criteria through a consideration of motivations. Thus motivations serve as a control strategy for selection.

- Selection is based upon both implicit and explicit criteria. The implicit criteria are used to make the potential space of revisions sensible and manageable by constraining the revision space. A complete account of selection must consider and specify both kinds of selection criterion.

This chapter addresses the problem of theory selection, of choosing one theory from amongst many based on an evaluation according to certain selection criteria. It begins with a broad survey of selection criteria and justifications for their use in a general context, and continues by noting some of the difficulties that arise. The next sections consider the different demands made of selection depending on the motivations of the reasoner, and how different kinds of selection are possible. Then a proposal is made for concrete computational implementations of appropriate selection criteria in the MID system, and for the mechanisms that manipulate them. Finally, related work is discussed.

## 8.2 Selection Criteria

In appraising what is commonly known as the scientific acceptability or credibility of a hypothesis, it is usual to consider the extent and character of the relevant evidence available in support of that hypothesis. (This has already been addressed in part in Chapter 6.) In addition, a variety of further criteria for theory acceptability have been proposed. Despite the debate over the use and value of different criteria [52], [79], the need for selection criteria in some form is generally accepted. McAllister [79] has divided selection criteria into two groups which he describes as indicators of *truth* and indicators of *beauty*. We shall not distinguish here between classes of selection criteria, for they are too strongly related to do so, but discuss them on equal merit. Quine and Ullian [88], for example, list a number of *virtues* of a hypothesis that make it more acceptable, regarding both truth and beauty in McAllister's terms. The following discussion introduces many of the general criteria that have been proposed for theory selection from a number of different perspectives. It draws examples from many sources including [88], and uses them to illustrate the use of selection criteria in both scientific and non-scientific contexts. It will subsequently be shown how these criteria can be accommodated in the MID system.

### 8.2.1 Accuracy

Accuracy is perhaps an obvious requirement of a theory, because the power to predict and explain are necessarily dependent on it. Typically, accuracy is an implicit consideration, ruling out any theories that do not submit to its requirements immediately, before any other judgements may be made. Accuracy was already mentioned in the previous chapter as being implicit, but is included here for completeness as there are varying degrees and forms. As Kuhn [52] points out, competing theories may display accuracy in different areas. He contrasts the oxygen theory which accounted for observed weight relations in chemical reactions with the phlogiston theory which accounted for metals being much more alike than the ones from which they were formed. The theories are incompatible, yet one matched experience better in one area, while the other was better in another. In such cases, a decision on the basis of accuracy would require a recognition of the area in which accuracy was more significant. By itself, accuracy is rarely a sufficient criterion for theory choice, for there may be innumerable accurate theories which explain the phenomenon under investigation.

### 8.2.2 Internal Consistency

A theory must not contain internal contradictions or inconsistencies. This is a logical requirement. If a theory is not internally consistent, then it will simultaneously predict mutually exclusive events as a consequence. McAllister [79] gives the example of the Aristotelian theory of free fall which asserted that heavier bodies fell faster than lighter ones. In envisaging a heavy body attached by a cord to a lighter one, Galileo was able to infer opposing conclusions: that the light body would slow the heavy one making the composite slower than the heavy body alone; and that since the composite was heavier than the heavy body alone, it should be quicker. Internal consistency is closely related to the *observation constraint* discussed in the previous chapter. In revising a theory, only those theories which are consistent with observations, and thus internally consistent, should be generated.

### 8.2.3 Historical Consistency

Historical consistency is closely related to accuracy, but it extends backwards over all previous experience, requiring that a theory is accurate for each instance. This, too, is an obvious requirement, but in the absence of a suitable theory, it may be relaxed. Our knowledge of the world changes over time, sometimes through a better understanding of the world, but more commonly through a changing reality. Depending on the kind of reasoning and the domain, we may be prepared to accept the possibility of historically inconsistent knowledge to a greater or lesser degree. Historical consistency, as will be shown later on in this chapter, is particularly significant.

### 8.2.4 Conservatism

It is quite likely that a new hypothesis or theory will conflict with prior knowledge or beliefs, but the fewer conflicts the better. If a theory conflicts with no prior beliefs then it is preferable since it is reconciled with what is already known. Conservatism is particularly useful, however, in dealing with irreconcilable differences.

It was, for example, thought that the planets revolved in circles around the sun. Tycho Brahe's observations, however, suggested otherwise, that their motion was not circular. Though the subsequent model of the planets was extremely revolutionary, it used the virtue of conservatism in that it retained the model of the sun being at the

focus, while altering the idea of circular planetary motion to elliptical planetary motion. Alternatively, consider the case of a magician telling us what card we have drawn from a pack. We might hypothesize that he is telepathic or is well-versed in ancient occult arts, but this conflicts dramatically with our prior beliefs. We might hypothesize that the magician used sleight-of-hand, and although this might conflict with our belief in our perceptiveness (since we didn't notice it), it is a far more conservative theory which is more plausible for being so. (In the field of mathematics, Imre Lakatos [57] illustrates conservatism very effectively in the form of a discussion between a teacher and a group of students.)

Conservatism sacrifices as little as possible of the evidential support that has been used to construct the knowledge that we have so far. Although this knowledge may be incorrect, conservatism forces a series of small revisions minimizing the error at each step, rather than a single large revision which may be entirely erroneous.

### 8.2.5 Simplicity

Simplicity is probably the most interesting of the criteria considered here, and has received most attention in the literature. It is also particularly difficult to characterize and justify. Simplicity is based on the principle of Occam's Razor which states that entities should not be multiplied beyond necessity. The best illustration of simplicity is in plotting points on a graph and then drawing a curve through them. Although there are an infinite number of curves that can be drawn, we will choose the simplest one which passes through or reasonably close to all the plotted points. The simplest curve, in geometric terms, is the one whose curvature changes most gradually from point to point. This represents a generalization that allows us to extrapolate through to untested points.

Simplicity demands context, however, and thus we may consider simplicity of a part compared to simplicity of the whole. Commonly, simplicity of the part is sacrificed for the greater unifying simplicity of the whole whenever possible. Consider Newton's hypothesis of universal gravitation stating that all bodies attract each other in direct proportion to mass, and inversely proportional to the square of the distance. In comparison to the intuitively simple hypothesis that heavy objects tend downwards, Newton's is far more complicated. It was, however, simpler in a greater sense, since it applied to a far wider range of phenomena, and allowed him to propose his unified system of terrestrial and celestial mechanics which had previously been explained by separate systems.

A major difficulty with the idea of simplicity lies in finding some way of measuring it. In the case of theories, it has been suggested that the number of basic assumptions or auxiliary statements might be used, but there is a problem in counting them. Consider the statement that for any two points there is exactly one straight line containing them. This might be counted as two statements — that there is at least one such line, and that there is at most one such line — rather than as one. Even if the count could be agreed, different basic assumptions might have different degrees of simplicity, requiring that they in turn would have to be evaluated. This problem with counting statements is one that detracts from many accounts of simplicity.

In itself, it is difficult to justify the merit of simplicity, since complicated theories are often formulated and used to accommodate new data. Yet just as conservatism was justified on the grounds that smaller leaps are better than larger ones, so simplicity is justified on the grounds of less complicated leaps. The more complex and intricate the hypothesis, the more ways of erring there must be, since there are more and wilder alternatives to choose between. Although the theory may need to be complicated, it is better to complicate the theory gradually, preferring the simplest theory possible at each step, in order to limit liability.

### 8.2.6 Generality

A hypothesis, if it is to be tested, must at least be sufficiently general to apply to more than just a single instance of the phenomenon under investigation. When we find a piece of copper wire that conducts electricity, we expect all copper to conduct electricity rather than just long thin copper wire. Generality is related to simplicity — it is simpler to believe that the observation that the copper wire conducts electricity is not a special case. The example given to illustrate simplicity will thus also serve for generality. If we take the two systems of terrestrial and celestial mechanics and compare them as a bipartite system to Newton's hypothesis of universal gravitation, then if the two taken together cover all that Newton's unified laws cover, there is no reason to prefer either on the grounds of generality. But the greater simplicity of Newton's system suggests a preference for the single system. Thus we may consider simplicity and generality together, complementing each other, and when we can maximize generality with little loss of simplicity, or gain simplicity with little loss of generality, then we have desirable properties in a theory.

### 8.2.7 Modesty

One hypothesis is more modest than another if it is weaker in a logical sense — if it is implied by the other without implying it. It contrasts with generality, and is motivated by different concerns. The hypothesis that birds fly with the exception of some such as penguins, for example, is more modest than the hypothesis that all birds fly. Modesty features in avoiding extravagant hypotheses which are relatively implausible in the normal course of events. It is closely related to conservatism, yet provides a basis for selection even when hypotheses are compatible with previous beliefs.

### 8.2.8 Refutability

If a theory is not refutable by any observation, then it is of no value in making predictions since these can never be tested. Popper claims that refutability is of paramount importance, and refuses even to consider any hypothesis that is not refutable or falsifiable [87]. An example of this is astrology. The vagueness of predictions and descriptions derived from the positions of the stars rule in too much as being possible, yet even if a prediction fails, it can be claimed that there is some item of information such as a planet's position at some time in the distant past that has been overlooked. Thus, any inadequacy is smoothed over, and conflict with other beliefs is avoided. This work does not share the same agenda as Popper; our concern with refutability is pragmatic. In the programme for reasoning described here entailing experimentation, refutation and revision, only those theories which are refutable can take part. Those which are not refutable are simply not susceptible to this kind of reasoning.

### 8.2.9 Confirmation and Corroboration

Many methods for suggesting greater or lesser degrees of support for a theory in the light of evidence have been suggested. This support, or *confirmation* of a hypothesis is traditionally considered to increase with the number of favourable test results. However, the increase in confirmation resulting from a single favourable instance will become smaller with an increased number of previous favourable instances. Thus if a large number of confirming test results already exists, one more confirmation will make little difference to the overall degree of confirmation.

In addition to the quantity of confirming instances, the diversity of confirming in-

stances is also important, since the greater the variety, the stronger the resulting support. Hempel [38] cites the example of Newton's theory of gravitation and motion which is supported by confirming instances from the experimental and observational findings for the laws for free fall, for the simple pendulum, for the motion of the moon about the earth and of the planets about the sun, for the orbits of comets and man-made satellites, for tidal phenomena, and so on. All of these laws were implied by Newton's theory, hence providing confirmational support for it.

Another dimension involves the stringency of the tests that provide the confirming instances. If a test can be made more precise through better experimental, observational and measurement procedures, then its results carry greater weight. Thus with greater number, variety and stringency of tests, the degree of confirmation accorded to a particular hypothesis increases.

Various systems for wrapping up the notion of confirmation in formal theories based on confirmation as probability have been proposed which make it possible to determine certain probabilities *provided that others are already known*[1]. The notion of confirmation, however is contentious. Without going into any detail, it should be noted that it admits paradox (for example, Goodman's paradox [32] and Hempel's paradox [37].) Furthermore, some philosophers deny that experience ever *confirms* hypotheses. Popper in particular proposes instead the notion of *corroboration* which is based on the idea of falsification rather than verification. According to Popper [87], a theory of high corroborability is one that is highly testable and of high content, and hence of low probability. The higher the content of the theory, the greater the opportunity to falsify it, and accordingly the lower the probability of it surviving. This illustrates the difficulty surrounding the issue. The use of evidence in defining a measure for the degree of corroboration of a theory is thus problematic.

## 8.3 Interaction and Overdetermination in Selection

There is much debate as to whether a complete list of criteria to be used in algorithms able to make an unequivocal choice exists, and whether it will ever exist. Two main problems arise: first, all of the individual criteria of choice must be unambiguously stated; and second, if such a thing is possible, then an appropriate weight function for the joint appli-

[1]Hempel's italics [38]. This will be discussed in more detail later on.

cation of the relevant criteria must be found. Kuhn [52], while accepting the possibility that notions of acceptance criteria may be broadly the same, rejects the possibility of a unanimous algorithm for theory choice by virtue of the subjective considerations which any individual uses to deal with these difficulties. These factors are said to be dependent on individual biography and personality. The significance of these claims to our work lies not in an historical critique, but in the recognition that there are external factors upon which the construction of an algorithm for theory choice and acceptance relies. Clearly, the criteria considered here for evaluating theories are strongly inter-related. In evaluating and selecting a theory, they must be considered together, as a whole, rather than as an amalgamation of separate parts. Generality, for example, is of little value if simplicity is sacrificed, and if there is no generality then simplicity is useless. In addition, modesty opposes the generality, and conservatism constrains simplicity, and they must be balanced against each other.

The use of a number of criteria in theory selection also leads to a different yet related problem. If the various criteria prefer alternative theories, then the problem of *overdetermination* arises where no unique theory can be selected [79]. This demands that some sort of conflict-resolution procedure be found, which can distinguish between rival choices. The problems of interaction and overdetermination are both serious ones which must be addressed, but which are often neglected in accounts of theory selection.

In considering these issues, Kuhn suggests that the criteria used be renamed *values*, since values provide effective guidance in the presence of conflict and equivocation without specifying the decisions to be taken. For example, freedom of speech is a value, but so is preservation of life and property. However, these two values often conflict so that a compromise between the two must be reached in order to prevent violent conflict yet allow progress. In the same way, the criteria used for theory selection are used as values rather than as rules. Values such as conservatism, modesty, simplicity and generality may prove ambiguous in application, but they do specify the factors influential in making a decision. Such an approach allows for the different aspects of scientific behaviour which may have been seen as irrational, and perhaps more importantly, acknowledges the different emphases placed on theory choice at different stages in the development of a theory. We shall see later how these *subjective* influences may be incorporated into a model of inductive discovery guiding the application of selection criteria so that the problems of interaction and overdetermination can be resolved.

## 8.4 Motivated Selection for Knowledge and Action

Existing systems that address theory selection are primarily designed for scientific discovery in some form (see section on related work). Their domains are, in the main, scientific domains — phlogiston theory, osmosis, etc — but if not scientific, they are still concerned with the discovery or evaluation of knowledge. ECHO [118], for example, has also been applied to cases of legal reasoning, but its motivations remain the same. These systems are geared to the acquisition of knowledge, not to the application of that knowledge. If we consider such systems to be autonomous reasoning agents, then their motivations are pre-determined and fixed, and very narrow. That is, the reasoning is designed to acquire knowledge rather than use knowledge. Because of this, there is a distinct bias to the kinds of revisions that are preferred in terms of the relative significance of the various selection criteria used. In particular, the merit of consistency, generality and simplicity is very important, and they are valued much more highly than modesty.

If we consider the motivations of a system that reasons on the basis of a need to take action, then we look at these criteria from an alternative point of view. The motivation for action implies an immediate need which must be satisfied in order to prevent the motivation from growing stronger, and a corresponding greater need to act.

In reasoning under knowledge motivations, the traditional path of maximizing consistency, simplicity and generality is followed. In reasoning under action motivations, however, efforts directed at increasing generality and simplicity are misspent, for the future applicability of the theory is not at issue. What is at issue is the ability of the system to reason correctly and adequately in the *current* situation only, to which modesty (and to some degree conservatism) are better suited. The primary consideration in reasoning under action motivations is precisely the need to take action. Concerns of historical consistency with previous episodes impose time and resource constraints which may be unacceptable. Historical consistency is important for it ensures that evidence already accumulated is used in the construction of new theories, and serves to rule out those revisions which cannot account for certain instances. However, the extra reasoning (processing) involved in such checks cannot be accommodated, particularly if the event histories are large, without violating the need to take action. Furthermore, in allowing the possibility of accepting poor quality evidence to be used in revision and selection, it is not unlikely that an inconsistency may result. There is little point in checking for

consistency with evidence that may, in different circumstances, be deemed inadmissible. Modesty limits the scope of the theory to the current scenario, thus ensuring the least possibility of error. Conservatism keeps the revision as close as possible to the original theory, minimizing the risk of faulty revision by keeping differences small. The need for action thus rules out a concern for generality and simplicity, for they increase the possibility of error in a specific situation. The need for action demands only that we take the correct action in order to satisfy our immediate needs whatever they may be (though we may subsequently generate a more generally applicable solution once the immediate need has passed, as we discuss later).

It could be argued that it is *urgency* that demands the different biases in selection rather than the kind of motivation. The consideration of reasoning for action presupposes the existence of a degree of *urgency*, however, while reasoning for knowledge excludes this. Though some needs may be less pressing than others, the desire for action necessarily implies a time limit of some sort. Thus urgency is implicit in reasoning for action.

The split between knowledge and action is, to some degree, contrived. In the discussion of motivations in Chapter 3, we noted the interaction between the two, and the need to address them together. There may be more subtle variations on the combination and use of selection criteria than these two, but the traditional uses for computational systems suggest only these immediately. Moreover, this distinction is widely held and is supported by a variety of research (eg. [56], [98], [63]).

Just as Kuhn points out the unknown subjective influences in the manipulation of his criteria as values in his historical (and philosophical) approach, so too such factors should be considered in a computational approach that seeks to provide a general mechanism for theory selection. In contrast to Kuhn, however, we assert that these influences may be explicitly modelled as motivations, with the knowledge-action distinction being seen as a top level characterization. A set of criteria or values for theory choice must be balanced in a way specified by the motivations of the reasoning agent. A limited model of motivations has already been developed in Chapter 3, and it now remains to develop a mechanism for allowing these motivations to serve as a control strategy for selection.

### 8.4.1 The Modification of a Domain Theory under Motivation

The use of different selection criteria depending on whether the motivations are biased towards knowledge or action implies that different domain theories will be constructed

as a consequence. But if a theory is to be modified, and if it is to be of subsequent use in reasoning about the world, then that modification must be one that is effected under motivations for knowledge.

If the intention is to take some particular action, however, so that modification occurs under motivations to accomplish tasks, then a theory will be constructed which is designed to apply specifically to the goal that action is designed to satisfy. In other words, the new theory will be valid for exactly the same kinds of situation as that in which the modification occurred, but will not necessarily be valid for any other. This contrasts with the case of reasoning for knowledge when the revised theory must be applicable to all other instances.

In between the two extremes lies a range of other situations in which motivations may be differently balanced. Accordingly, a whole series of different theories may be constructed in the course of reasoning under these different motivations. In situations in which the primary motivation is not that of knowledge, the benefit that may be gained from reasoning inductively is destroyed, since for each situation encountered, a new domain theory will need to be constructed, tailored to the requirements of that particular situation. This serves no purpose, for not only does such a system not learn from experience, it also requires far greater use of resources by reasoning in each situation. Furthermore, the lack of a sufficiently general theory from which to derive more specific theories may negate the possibility of generating those specific theories, and the only justification for this approach is compromised. On this premise, the merit of the entire programme proposed here is in doubt. Clearly, this is an unreasonable way to proceed. Below, we propose a method of temporary revision to address this.

### 8.4.2 Permanent and Temporary Revision

When a system reasons with the aim of inductively acquiring knowledge, it attempts to construct sufficiently general theories that may be used subsequently. Reasoning with the aim of taking action requires a more specialized domain theory. This specialized theory, however, is only appropriate for the immediate circumstances, first because of the bias imposed, and second because of the possibility of poor quality evidence as we have noted above, and is of no use in subsequent situations. Thus we may consider it to be a temporary theory, derived from the more general theory that is applicable in other situations, and discarded when it is no longer of use. Consequently, we have two different

kinds of revision which occur at different times, depending on the motivation.

Reasoning to acquire knowledge permanently revises a theory to produce ever more general and accurate theories. It is on the basis of these theories that we are able to reason effectively and learn from experience. Moreover, it is on the basis of these theories that we are able to derive the temporary theories which apply to specific situations.

Reasoning to take action, on the other hand, revises a domain theory when the general theory is inadequate for the particular task at hand, but it just produces a temporary theory which is discarded after use (or after the current situation has passed).

As was pointed out in Chapter 3, however, even when we reason under action motivations, we are still able to add to our knowledge. New knowledge that is discovered in the course of reasoning for action may also be used in a subsequent permanent revision. Thus temporary revision is only one part of the revision process, for there may also be an associated permanent revision with a different bias that attempts to incorporate new knowledge into the underlying general theory. (This is possible only when the evidence is of sufficient quality to be used for a permanent revision, however.) In this way, the domain theory retained is a global general version, with local specialized versions being constructed from it at appropriate times. This satisfies the need for generality by default, and also the need for localization when confronted with situations in which the general theory is inadequate, but avoids the problem of maintaining multiple domain theories.

Although it might be argued that there are finer distinctions to be made than that simply between knowledge and action, this still holds. The different kinds of motivations that lead to inductive reasoning can be regarded as a hierarchy at the top of which lie the twin motivations of knowledge and action. Lower levels, while differing from the higher levels, are more refined instances which are, nevertheless, subsumed in the higher levels.

## 8.5 Selection in MID

### 8.5.1 Discovery and Justification

The distinction between discovery and justification has been widely recognised, but is increasingly being disputed (eg. [138], [58]). Discovery is concerned with the formation of hypotheses to account for observations and to make new predictions, while justification is concerned with the issues of theory selection and acceptance. The use of selection criteria is usually confined to the context of justification in both philosophical computational

accounts. The potential use of selection criteria in the context of discovery, however, has been noted. Achinstein [1], for example, suggests that selection criteria are also suitable for applying constraints to the kinds of hypotheses that are proposed in the first instance. Their use is most apparent when considering revisions to hypotheses, since the hypotheses are created from a base hypothesis, albeit an incorrect one.

We assert that the use of selection criteria in the context of discovery or in the generation (revision) of theories is not only possible or desirable, but necessary. It is not possible to generate every possible revision that may be consistent with the observations, particularly with large and complicated theories. Any attempt to do so would lead to the problems of combinatorial explosion. Even if it was possible to generate every possible revision, this would take an inordinate amount of time, in both generation and discrimination, so that in cases with action motivations the reasoning agent may not be well served. It is necessary therefore, to have some means for providing constraints on the *revision* space through the use of selection criteria [83].

Selection criteria may thus be used in two ways. First, they constrain the kinds of revisions that are generated, or at least order the revisions so that better ones are ranked higher and considered earlier. Second, given a set of competing theories, selection criteria provide a basis for discriminating between them. In both cases, application of the criteria is guided by the biases produced from a consideration of the relevant motivations.

### 8.5.2 Overview

We can define selection criteria in terms of preferences for revision operators which constrain the generation of revisions corresponding to the context of discovery, and in terms of judgements on candidate revisions through the numbers of components once they have been generated, corresponding to the context of justification. We refer to these as *dynamic* and *static* selection criteria respectively.

Figure 8.1 shows the two levels of selection in MID. At the top, the dynamic selection determines the order in which the different revision operators should be applied. An operator may produce more than one revision, in which case these must be further discriminated by the second *static* selection mechanism which evaluates candidate theories after revision. Thus *dynamic* selection imposes a rough ordering on the revision space, and *static* selection refines this by ordering subclasses of revision. Each revision that is generated is checked, in order, for historical consistency to the required level. If it is not

consistent, then the next revised theory is checked, until the current subset of revisions is exhausted, at which point the next revision operator is applied to generate a new subset of revisions. When a consistent theory is found, it is accepted, and no further search of the revision space is necessary. The thick black arrow indicates the direction of search through the space of revisions. The thin arrows indicate the revisions after they have been checked for consistency with prior episodes. Those with dashed lines are rejected as inconsistent, while the longer solid arrow represents the final consistent theory.

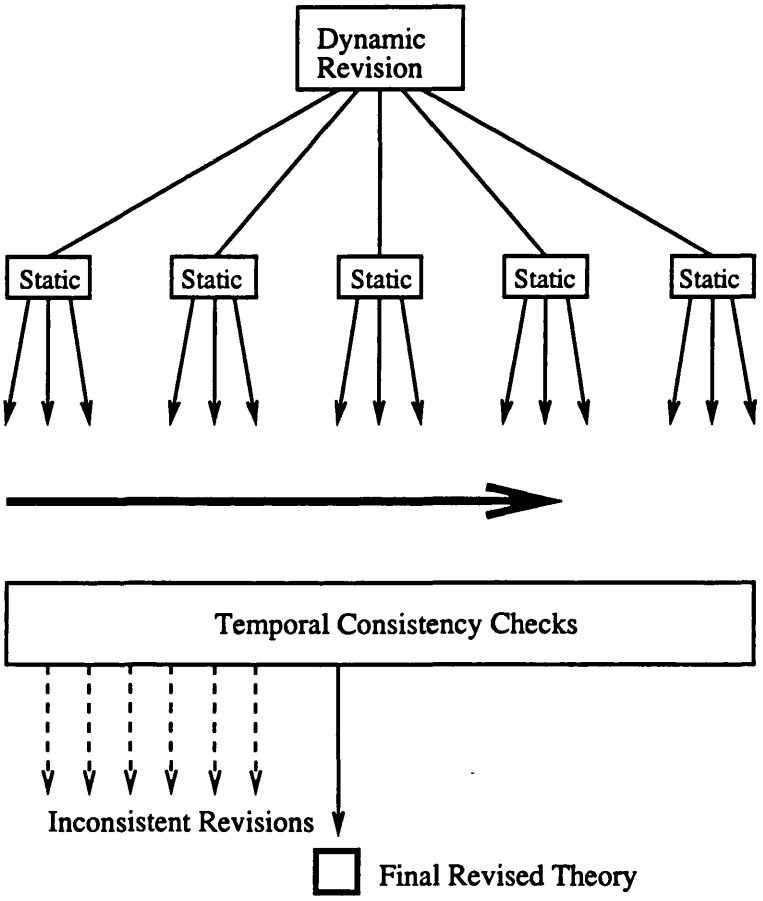



Figure 8.1: Dynamic and Static Selection in MID

The selection stage in MID is summarized in the Table 8.1. It is discussed in detail over the course of the following sections.

### 8.5.3 Specification of Selection Criteria

In discussions of selection criteria, it is common for the different concepts to be very vague and not well defined. This is partly due to the difficulty of adequately capturing a precise

**Input:** $X$, a list of intermediate revisions initialized with the original domain theory, $\Delta$, and the anomalies arising.
**Output:** a revised theory $\Delta'$
**Algorithm:**

- Rank the revision operators according to the values determined by motivations (dynamic selection)
- Get the next state from $X$
  - If more anomalies remain, get the next anomaly.
    * Get the next unused operator applicable to the current anomaly.
    * Generate the next revision(s) state.
    * Order the resulting revision(s) according to the values determined by the motivations (static selection).
    * Add the new (ordered) revisions to $X$.
  - If no anomalies remain in the current state, then the revision is completed.
    * Check the consistency of the completed revision.
    * If the revision is consistent with past experience:
      – return first completed revised theory
      – if it is a permanent revision (motivated for knowledge) update the domain theory to get $\Delta'$
    * If the revision is not consistent, then remove the current state from $X$

Table 8.1: The algorithm for selection in MID.

definition that is generally applicable. In work on artificial intelligence where knowledge must be explicitly represented in some language, this is even more of a problem. Any specification of criteria that is tied to a particular knowledge representation cannot but suffer from a lack of expressiveness. Nevertheless, this must be done, and an *operational* specification of the individual criteria for Qualitative Process Theory is developed below.

### Accuracy

Accuracy is an important requirement of this system. The space of revised theories is constrained by the observations that are provided as input so that all are consistent. It is an implicit constraint on revision. In part, accuracy was considered previously in the evaluation stage.

### Refutability

Refutability, too, is implicit, and is demanded by pragmatic considerations. Since this is an empirical programme for reasoning which proceeds by conjecture and refutation, irrefutable theories cannot be considered. In adopting this requirement on theories, the concern is not with what is or is not science, but what can be effectively reasoned about. The fixed representation of theories enforces this constraint, and it does not play a role in the selection of alternative theories.

### Conservatism

Conservatism involves the least conflict with prior beliefs. It suggests scope changes, retaining the underlying structure, but making a process more accurate. It also prefers adding components, retaining beliefs from previous observations, suggesting that some things may have been missed, but that no observations that were made were incorrect.

**Dynamic** Operators that have positive effects on conservatism are: Specialize- Component, Generalize-Component, Add-Component. Operators making a negative contribution are: Negate-Component, New-Process, Delete-Component.

**Static** Conservatism prefers the smallest possible change in numbers of components.

### Simplicity

Simplicity is related to the structure of a theory, but also to the degree of unification. It prefers fewer processes, unifying the theory into a coherent whole, and fewer conditions and disjunctions of conditions, simplifying the structure of the theory. It serves to avoid the excesses that may arise through generality.

**Dynamic** Operators that have a positive effect on simplicity are: Delete-Condition, Specialize-Condition. Operators that have negative effects are: New-Process, Add-Condition, Generalize-Condition.

**Static** Simplicity prefers minimizing the number of processes, the number of conditions and the number of disjunctions of conditions.

### Generality

Generality is vital if we are to be able to predict future events. A more general theory is one which entails more predictions. In part, generality is demanded by the scenario and observation constraints on revision, and the consistency criterion here. These require that revisions are consistent with the current evidence and, as far as possible, consistent with previous observations. There is still scope for preferences on operators, however.

**Dynamic** Operators having a positive effect on generality are: Generalize-Component, Delete-Condition, Add-Effect. Operators that make a negative contribution are: Specialize-Component, Add-Condition, Delete-Effect.

**Static** Generality prefers minimizing the number of conditions and maximizing the number of disjunctions of conditions, and effects.

### Modesty

With motivations to take action, modesty is used to find a more specific theory. It is the exact opposite of generality. Thus it prefers revisions which entail smaller applicability.

**Dynamic** Operators that make a positive contribution to modesty are: Add-Condition, Specialize-Component, Delete-Effect. Operators that make a negative contribution are: Generalize-Component, Delete-Condition, Add-Effect.

**Static** Modesty prefers maximizing the number of conditions and minimizing the number of disjunctions of conditions, and effects.

### Evidential Support

Due to the difficulties with confirmation and corroboration, a naive notion of evidential support is used, simply considering the time since the last revision of each process. (This is a rough heuristic bearing some similarity to the *confidence measures* of KEKADA [53] noted in Chapter 6, which direct attention towards more promising possibilities for exploration.) Processes that have received more support than others (have not been revised recently) are preferred for revision over processes with less support (have been revised recently). This involves maintaining a revision history for each process in the domain. It should be apparent that a measure of evidential support relies upon comparison. There is no obvious way of using this notion in selecting operators for revision independently,

only in evaluating the domain theory first and then using operators appropriately. This, however, negates the advantage to be gained by the two step selection process, since it requires both a pre-evaluation and a post-evaluation of domain theories. Consequently, the evidential support criterion is used only in the second, *static* step of evaluating candidate revisions. Evidential support is used to direct attention towards theories which are more likely to contain errors through recent revision.

### Historical Consistency

The requirement of historical consistency enforces consistency with previous scenarios and observations. It is needed to avoid *oscillation* in revision by which a revision is made on the basis of one set of observations, but is retracted through a subsequent set of observations with inverse effects. It may happen that a domain theory $\Delta_1$ is revised as a result of a set of observations to produce $\Delta_2$. After more observations, the domain theory may subsequently be revised, but this time back to the original $\Delta_1$. Checking consistency with previous cases would avoid such problems. This requires that a history of previous scenarios and observations is maintained so that revisions which are selected on the basis of other criteria can be subjected to a final check on the basis of historical consistency. The historical consistency criterion addresses the *experience consistency problem* which has been treated as a separate stage in discovery by other work [90]. However, the requirement on historical consistency *is* a selection criterion, and is accorded the same status in this model. Furthermore, it is subject to the same variation in importance as other criteria, and can be overridden. As stated earlier, knowledge is not static in that our understanding of the world changes, and the world itself changes. An awareness of this gives licence to relaxing the historical consistency requirement, since a theory need not be consistent with experience when that experience is redundant.

In MID, historical consistency is achieved by rejecting revisions which are not consistent with the history of previous events. This creates a substantial overhead in storage and computation, since each revised theory must be used to generate predictions which are compared against the cases in the history. There are, however, ways to minimize this overhead. First, we can limit the history to the $n$ most recent scenarios, where $n$ is a system parameter. Second, in the case of reasoning under knowledge motivations, if no candidate revisions are consistent with the history of events, then it may be desirable to discard a set of scenarios from the maintained history. These scenarios would be only

those prior to scenarios which are consistent with the current scenario. This is justified if the state of the world is known to change such that certain results may become redundant at some point in the future. Third, we can try to compress the number of scenarios in the history through various techniques such as the use of exemplars [90]. A fourth possibility is to maintain a record of the critical points in the development of the theory — only those episodes that involved revisions — but this assumes that the theory will always be built up from nothing. This assumption will not always be valid, for we want to be able to reason in cases where a possibly faulty domain theory already exists. In such cases, it is also important to maintain records of episodes that do not involve revisions.

Currently, MID implements the first strategy, keeping a limited number of scenarios in the history, for the sake of simplicity. This is also augmented with the second strategy, using information about changes in the world to discard redundant scenarios.

The assumption implicit in all of this is that changes in the world are static rather than continuous. That is to say that certain periodic phenomena lie beyond the abilities of the existing system. In considering tides, for example, repeated observations will give evidence of the tide being in, say. Now when the tide changes, the system notes an inconsistency with previous instances in the history and must resolve it. (Although multiple processes in the QPT formalism can be used to represent this behaviour by having one process to represent the tide coming in and another to represent the tide going out, for example [22], there are limits to the expressiveness of this.) The available options ignore the possibility of continual periodic changes, and offer ways only of eliminating the inconsistency. Thus a restricted view of the world is assumed. MID could, however, be enhanced to incorporate ways of reasoning about such periodic phenomena by including temporal information. At points of disagreement between observations and the history, the time difference could be used to hypothesize possible periodic relationships which could then be tested against subsequent observation. This would, however, place even greater demands on the history of events in terms of space requirements in order to maintain sufficient details to entertain periodicity hypotheses.

### 8.5.4 Combining Selection Criteria

To recap on the model of motivations described in Chapter 3, we have, for any reasoner, a set of motivations, $M$, each motivation being an *m-triple* $< m, v, b >$. The set of motivations can be broadly divided into knowledge motivations and action motivations.

As we have seen, knowledge motivations provide the impetus to acquiring knowledge while action motivations demand the taking of actions for particular needs.

According to this model, the motivation with the highest strength value, the ***salient motivation***, determines the nature of the reasoning. Normally, this will be the knowledge motivation, because it is fixed at a level that reflects our desire for knowledge, while the other motivations, the action motivations, are not as high. However, in certain circumstances, a particular action motivation will surpass the knowledge motivation and become salient. This indicates a greater importance, and a greater need to satisfy that motivation, resulting in a switch from the broad demands of acquiring knowledge, to specific focussed task requirements. In this way, the level of strength of the knowledge motivation acts as a threshold above which action is demanded. Thus, just as the level of the salient motivation determines the importance and the need to accept evidence, so it determines the relative importance of the selection criteria used in judging theories. A high salient motivation demands a strongly focussed theory through minimal use of resources, while a lower motivation, relaxes this requirement.

As discussed earlier, in reasoning under knowledge motivations, the traditional path of maximizing consistency, simplicity and generality is followed. In reasoning under action motivations, however, concerns with increasing generality and simplicity are replaced with concerns of modesty and conservatism to ensure that the system reasons correctly and adequately in the *current* situation only. This is related to the need for adequate evidence. If information is necessary in order to act, and if action is required urgently, then the acceptance threshold of evidence is reduced, allowing action on the basis of inadequate evidence. Poor quality evidence accepted in this way must not, however, be used to construct a permanent revision. Moreover, when the desire for action is strong, available resources may be limited leading to inadequate consistency checking, and resulting in a sub-optimal solution. Revising knowledge in accordance with this (though it may be acceptable in the circumstances,) demands that the revision (even a temporary one) be kept as local and specific as possible, and close to what was ***known*** previously.

### Weightings

As has been consistently emphasised, the weightings on the various selection criteria are determined by the underlying motivations of the reasoner. They are considered to be

*values* rather than weights because they are not devoid of meaning, but embody particular motivations. Broadly, the two kinds of motivations for reasoning can be associated with two specific value schemes: one that highly values consistency, generality and simplicity, the other that values conservatism and modesty.

It seems that further variations are possible. If the salient motivation is close to the knowledge motivation, then the relative importance of acquiring knowledge is still high, and can be incorporated into the values of the selection criteria. As the action motivation increases and the gap between the two grows, the values move towards the absolute values for action accordingly. In this case, we need to state explicitly the manner in which the criteria relate to each other through rules or equations:

$$\omega_{modesty} = (1 - \omega_{generality})$$

where $\omega_{criterion}$ is the value weight asociated with that criterion. Thus as modesty is valued more, so generality is valued less, and vice versa. Similar relationships can also be specified, and the strength of the motivation used to determine a minimal set of weightings with the remainder being derived from that set through these equations.

However, this admits a serious flaw. If action motivations are salient, by however small a margin over knowledge motivations, then some distortion on the revision that would be preferred by knowledge motivations alone must occur. In this case, there must be a second revision later anyway to avoid that distortion being carried through permanently. Consequently, any compromise that is made between the two extreme cases is redundant. There is no point in compromising action motivations by balancing weaker knowledge motivations, since this will just result in a less directed revision. There is no point in compromising knowledge motivations, since any compromise will demand a second revision making no concessions to action motivations. Apparently, a two-way split in motivations is necessary, at least at the level being considered here.

Therefore, MID uses such a split between values, having one set for knowledge motivations and another for action. Its values for knowledge, $\mathcal{W}_{knowledge}$, and action, $\mathcal{W}_{action}$, where $\mathcal{W} = < \omega_{conservatism}, \omega_{simplicity}, \omega_{generality}, \omega_{modesty}, \omega_{support}, \omega_{consistency} >$, are:

$$\begin{aligned} \mathcal{W}_{knowledge} &= < 0.8, 0.8, 0.8, 0.2, 0.1, 1.0 > \\ \mathcal{W}_{action} &= < 0.4, 0.2, 0.1, 0.9, 0.1, 0.0 > \end{aligned}$$

The earlier equation relating modesty to generality is used here, too, since the opposing nature of the criteria cannot be accommodated otherwise. These are defaults which

are determined based on the motivations of the reasoner. In the case of knowledge motivation, these will be high for simplicity, generality and consistency, while with action motivations, modesty is valued much higher and conservatism relatively higher, and the traditional virtues are valued less.

The values themselves are meaningful in imposing an ordering on the space of revisions. A value of 0 denotes no significance to the ordering, while a value of 1 denotes utmost significance. The values in between these limits denote greater or lesser significance. Thus a value of 0.4 for modesty means only that it has lower significance than 0.5, and greater significance than 0.3.

The values specified above are estimated values in the prototype ARA that MID represents. They are not necessarily the best values, but have been chosen to represent the biases imposed by the relevant motivations as discussed above. The justification for their use lies in the aim of MID to achieve certain results. Note that for the consistency criterion, the value is either 1 or 0 implying that consistency is switched on and off under different motivations. Potential variations in which degrees of consistency are specified are possible, but it is not clear how selective consistency checking might be done. The decision to have a binary switch was made for the sake of simplicity.

It is recognised that further investigation of the relationship between the motivations of an ARA and these value sets is appropriate.

### 8.5.5 Dynamic Selection

In order to avoid the problems of intractability discussed earlier, we must provide a means of using these criteria to guide the application of revision operators. We can do this through an informed depth-first search with backtracking using the selection criteria as heuristics on revision operators themselves [85]. This requires that each revision operator has associated with it some indication of merit as determined by the appropriate criteria. In MID, the operators can be pre-evaluated in terms of the contribution they make to each criterion, and an appropriate weight function used. As the relative importance of the various selection criteria changes with changes in motivations, so the weight function must change.

| *revision operator* | *selection vector* |
|---|---|
| Add-Condition | $< 1, -1, -1, 1 >$ |
| Delete-Condition | $< -1, 1, 1, -1 >$ |
| Negate-Condition | $< -1, 0, 0, 0 >$ |
| Specialize-Condition | $< 1, 1, -1, 1 >$ |
| Generalize-Condition | $< 1, -1, 1, -1 >$ |
| Add-Effect | $< 1, 0, 1, -1 >$ |
| Delete-Effect | $< -1, 0, -1, 1 >$ |
| Invert Effect | $< -1, 0, 0, 0 >$ |
| Specialize-Effect | $< 1, 0, -1, 1 >$ |
| Generalize Effect | $< 1, 0, 1, -1 >$ |
| New-Process | $< -1, -1, 0, 1 >$ |

Table 8.2: Selection vectors for all classes of revision operator.

### Selection Vectors

The evaluation of the contribution made by individual revision operators to each criterion can be found in the operational specification of the criteria above. We define a criterion in terms of the revisions that it prefers, and by extension in terms of the revision operators that it prefers. Thus we can define a ***selection vector*** for each operator, where the selection vector, $S_{operator}$, is an array of integer values, one for each selection criterion, positive, negative or zero. Positive and negative values indicate positive and negative contributions, while zero indicates no effect. Note that only four of the criteria specified earlier are suitable for use in this *dynamic* selection.

$$S_{\text{operator}} = < C_{\text{conservatism}}, C_{\text{simplicity}}, C_{\text{generality}}, C_{\text{modesty}} >$$

where $C_{criterion}$ is the contribution made to the criterion.

Given the weights and the selection vectors, it is possible to determine a partial ordering on the revision operators simply by multiplying the two together. The ordering is only partial, because we may not be able to choose between two or more operators. This requires the second stage to select revisions after generation, making use of the complementary static criteria as well as consistency and evidential support.

Table 8.2 gives a selection vector for all revision operators. There is no need to distinguish between the various conditions or effects, because they are treated as classes.

| | | |
|---|---|---|
| `flow-aligned` | → | `aligned` |
| `fluid-flow-aligned` | → | `flow-aligned` |
| `heat-flow-aligned` | → | `flow-aligned` |

Figure 8.2: An example background knowledge rule base.

### 8.5.6 Static Selection

The use of selection criteria in justification is different. It evaluates candidate revisions once they have been generated, and is centred around the structure of the theories through counts of processes and components in processes.

#### Granularity

This, however, leads to the problem that was identified in the original discussion of simplicity: the granularity of individual predicates is unknown. If predicates are abstracted to different levels, then it is not clear how we can count the number of components in a process, for some predicates may entail multiple predicates at a finer granularity.

Using the background knowledge, however, we can reduce the different levels of abstraction to the same one, and can then proceed to count components in the desired manner. Consider again the background knowledge rule base (BKRB) which represents knowledge about abstraction hierarchies in the form of implication rules (eg. Figure 8.2). Higher levels of the hierarchy involve more disjunctions of predicates than lower levels. By tracing through the rules, an approximation of the granularity of a predicate can be estimated. Suppose for example that `flow-aligned` is a condition in a process. If we trace a chain from `flow-aligned` to its end at `fluid-flow-aligned`, we can use the links in the chain to estimate the number of disjunctions. The end of the chain must be observable, and is thus primitive. Consequently, we can weight the condition with granularity giving, in the case of `fluid-flow`, a weight of 2 (the number of links + 1). Similarly, if the predicate was `aligned`, then the weight would be 3.

This allows a static structural evaluation of candidate theories based on the counting of components which is not compromised by the use of varying levels of abstraction in representing predicates. Thus we avoid the unstated and mistaken, though prevalent, assumption that all predicates have the same granularity. This method does require, however, that all abstractions are grounded in primitive predicates.

### Comparison of Candidate Theories

A static structural evaluation also allows the introduction of new theories that have not resulted from the revision process. Theories that are generated by other reasoning techniques (such as analogy) can enter the system here and be compared with existing theories on the basis of the static structural evaluation. Since they take no part in revision, they cannot be selected through the constraints on revision operators. We can define an evaluation selection function on a theory, $\Delta$, as follows:

$$f(\Delta) = \sum_{c \in criteria} \omega_c \times n_c$$

where $\omega_c$ is the *value* weight associated with the criterion $c$, and $n_c$ is the score of the criterion $c$ defined below.

The selection criteria for this second part of selection are also defined operationally. That is, the criteria are described by their specifications for counting components above. Specifically, we define the criteria according to Table 8.3. The selection function above can be used very easily by adding together the relevant counts for each criterion and multiplying the result by the value weight. An alternative selection function might exclude the evidential support criterion and use it subsequently. That is, the other criteria would be used as above, and then if no unique choice could be made, groups of theories could be assessed by using evidential support to impose a further sub-ordering. This avoids undue weight being given to evidential support in cases where competing theories have radically different times since the last revision. The disadvantage of this approach is that evidential support is not balanced against the other criteria in an even way. Moreover, the first approach allows a simple and elegant combination mechanism. (Note that although consistency is included as a static selection criterion in the table, it is not used in the evaluation selection function, but is used separately. Consistency is discussed in the next section.)

In fact, revisions are not scored in isolation according to the various counts of their components, but in relation to each other. The counts of the different components may be of varying magnitudes, and the product of these would not provide an effective evaluation. Instead, the counts that are generated for each revision are compared against the limit values for the set of revisions being compared, and the difference between particular revisions and the limiting revision for each criterion is used instead of the original count. Thus when preferring smaller counts, if one of the generated revisions had a lower compo-

| *criterion* | *count* | *preference* |
|---|---|---|
| Conservatism | change in no. of components | minimize |
| Simplicity | no. of processes | minimize |
| | no. of conditions | minimize |
| | no. of disjunctions of conditions | minimize |
| Generality | no. of conditions | minimize |
| | no. of disjunctions of conditions | maximize |
| | no. of effects | maximize |
| Modesty | no. of conditions | maximize |
| | no. of disjunctions of conditions | minimize |
| | no. of effects | minimize |
| Support | time since last revision | maximize |
| Consistency | no. of inconsistent prior episodes | minimize |

Table 8.3: Specification of static selection criteria.

nent count than all of the others, then the scores used in the selection evaluation function for these component counts would be the difference between this minimum value and the component counts in all the other theories. Similarly when preferring greater numbers of components, the difference between the highest count for a component and the counts for other theories is used. In this way, by using differences rather than absolute values, the selection function provides a more balanced combination of the selection criteria. An alternative would be to use proportions rather than differences between the particular and limiting counts. However, with very large theories and small changes, all of these ratios would be similar, and the weighted products would be dominated by the weight values rather than a more equal division between the weights and the counts themselves. Using the difference instead ensures that this division of responsibility between the weights and counts is kept constant even with very large numbers.

### 8.5.7 Consistency

Each revision that is generated is checked, in order, for historical consistency to the required level (which may be minimal in the case of reasoning for action)[2]. If it is not consistent, then the next revised theory is checked, until the current subset of revisions is exhausted, at which point the next revision operator is applied to generate a new subset of revisions. When a consistent theory is found, it is accepted, and no further search of

[2]MID imposes an all or nothing requirement on consistency, so that a revision must either be entirely consistent, or need not be at all consistent.

the revision space is necessary.

If, after checking each possible revision against the maintained history of previous instances, no consistent revision remains, then there are a number of alternative strategies that can be pursued. Some of these have already been mentioned. It is important to note that these cases of inconsistency will be rare, and will arise through a serious problem, either an error in the learning process to date, or a fundamental change in the world that makes some previous results invalid. This will therefore require some form of exceptional repair to the current knowledge structures.

If we assume that the state of the world changes such that certain results may become redundant at some point in the future, then the history may be modified to allow a consistent revision. This is done by discarding those instances in the history before a particular episode that prevented the generation of a consistent theory. The difficulty here is in deciding when such an action is warranted. It might be achieved by replicating the conditions of the episodes in question, and repeating the experiments to determine whether the results are still valid. The disadvantage of this approach is that it requires substantial extra experimentation. If there were an easier way to do this, then it would be an attractive option, because it would also limit the number of instances in the history, with good justification for doing so, rather than in an arbitrary way.

Another option is that of modifying background knowledge. Previously, the possibility of revising background knowledge was mentioned. However, if a consistent revision can be generated, then there is no reason to do so, for background knowledge is held with a greater certainty than the domain theory itself, and is not the main subject of consideration. An inconsistency, however, indicates a serious problem in the existing knowledge that goes beyond the domain theory itself, and merits investigating the possibility of whether a revision to the background knowledge might be an appropriate solution.

Finally, the current episode may be rejected. If an inconsistency arises, it might be due to problems with the experimental scenario and observations that caused it.

The occurrence of an inconsistency directs attention away from the default revision mechanism that applies to the domain theory but which cannot account for the observed anomaly. Instead, attention is focussed on the other repositories of knowledge that are used by the system. Each of the three alternatives mentioned here considers revision to a different repository of knowledge. However, it is not clear how a choice could or should be made between these alternatives in assigning blame for the inconsistency. MID

```
Process Name:          heat-flow
Individuals:           object ?source
                       object ?destination
                       heat-path ?path
Preconditions:         aligned ?path
QuantityConditions:    greater-than (a (temperature ?source))
                                    (a (temperature ?destination))
Relations:             Q+ heat-flow-rate (temperature ?source)
                       Q- heat-flow-rate (temperature ?destination)
Influences:            I+ (heat ?destination) (a (heat-flow-rate))
```

Figure 8.3: Another erroneous domain theory concerning heat flow

```
Scenario Name:   heat-flow-example2-scenario
Individuals:     std-object1
                 std-object2
                 std-path1
Facts:           object std-object1
                 object std-object2
                 heat-path std-path1
                 aligned std-path1
                 greater-than (a (temperature std-object1))
                              (a (temperature std-object2))
```

Figure 8.4: A scenario description in which heat flow occurs

implements a form of the first strategy, discarding currently inconsistent episodes from the history. First, MID assumes that the world is capable of change. Second, there is a practical benefit to be gained in terms of time and space requirements from a smaller history. Third, it seems to offer a compromise between the second option of making a radical change to the background knowledge, and the third option of assuming that the inconsistency is due to invalid evidence.

## 8.6 A Simple Example

### 8.6.1 Dynamic Selection

To illustrate the selection procedure, consider the domain theory specified by Figure 8.3. The theory contains knowledge about only one *process*, heat-flow. If the conditions are satisfied, then the theory predicts that the temperature of the destination object will increase. (The theory is erroneous in that it does not know about the requirement of a heat-connection between the objects involved in the heat-flow process. For the sake of clarity and simplicity, the error of the missing effect of heat flow, that the temperature of the source will decrease, will be ignored in this example.) Now, suppose MID is provided

```
Predictions:  increase (heat std-object2)

I+ (heat std-object2) (a (heat-flow-rate))
 Active Process:  heat-flow std-object1 std-object2 std-path1
                  aligned std-path1
                  greater-than (a (temperature std-object1))
                               (a (temperature std-object2))
```

Figure 8.5: The predictions generated by MID

| *revision operator* | *selection vector* |
|---|---|
| Add-Condition | $< 1, -1, 0, 0 >$ |
| Negate-Condition | $< -1, 0, 0, 0 >$ |
| Delete-Effect | $< -1, 0, -1, 1 >$ |

Table 8.4: Selection vectors for example revision operators.

with the scenario description of Figure 8.4, then the heat-flow process will be active since all of the conditions are satisfied. Accordingly, MID makes the appropriate prediction shown in Figure 8.5.

However, there are no observations, and the theory must be revised so that it is consistent with the observations. MID identifies a number of potential revisions that come under the headings of *revise conditions* so that the erroneous process is no longer active, and *revise effects* so that the erroneous predictions are no longer made. (For the sake of simplicity in this example, it is assumed that no background knowledge is available, limiting the possible revisions.)

Instances of the revisions that MID considers are: Delete-Influences, Add-Pconditions, Add-Qconditions, Negate-Pconditions, and Negate-Qconditions. The dynamic selection mechanism in MID orders these revision operators on the basis of motivations. In the case of reasoning under knowledge motivations, MID uses the value set $\mathcal{W}_{knowledge} =< 0.8, 0.8, 0.8, 0.2, 0.1, 1.0 >$. Given the selection vectors specified earlier, a subset of which are shown in Table 8.4, MID calculates an ordering on these operators using the first four values for conservatism, simplicity, generality and modesty, as follows:

$$\textbf{Add} - \textbf{Condition} \quad (1 \times 0.8) + (-1 \times 0.8) + (0 \times 0.8) + (0 \times 0.2) = 0.0$$

$$\textbf{Negate} - \textbf{Condition} \quad (-1 \times 0.8) + (0 \times 0.8) + (0 \times 0.8) + (0 \times 0.2) = -0.8$$

$$\textbf{Delete} - \textbf{Effect} \quad (-1 \times 0.8) + (0 \times 0.8) + (-1 \times 0.8) + (1 \times 0.2) = -1.4$$

Thus MID carries out the Add-Condition revision first, followed by Negate-Condition, and then Delete-Effect. Under action motivations,

| | |
|---|---|
| PROCESS: | heat-flow |
| Variables: | ?source |
| | ?destination |
| | ?path |
| Individuals: | object ?source |
| | object ?destination |
| | heat-path ?path |
| Pconditions: | aligned ?path |
| | precondition-2 ?source ?destination ?path |
| Qconditions: | greater-than (a (temperature ?source)) (a (temperature ?destination)) |
| Relations: | Q+ (heat-flow-rate) (temperature ?source) |
| | Q- (heat-flow-rate) (temperature ?destination) |
| Influences: | I+ (heat ?destination) (a (heat-flow-rate)) |
| Revision Log: | 1 add preconditions |

Figure 8.6: The revised domain

| | | |
|---|---|---|
| flow-aligned | → | aligned |
| heat-flow-aligned | → | flow-aligned |

Figure 8.7: A background knowledge rule base for the static selection example.

$\mathcal{W}_{action} =< 0.4, 0.2, 0.1, 0.9, 0.1, 0.0 >$, and we get:

$$\textbf{Add} - \textbf{Condition} \quad (1 \times 0.4) + (-1 \times 0.2) + (0 \times 0.1) + (0 \times 0.9) = 0.2$$

$$\textbf{Negate} - \textbf{Condition} \quad (-1 \times 0.4) + (0 \times 0.2) + (0 \times 0.1) + (0 \times 0.9) = -0.4$$

$$\textbf{Delete} - \textbf{Effect} \quad (-1 \times 0.4) + (0 \times 0.2) + (-1 \times 0.1) + (1 \times 0.9) = 0.4$$

In this case, MID carries out Delete-Effect first, then Add-Condition, and then Negate-Condition. Thus MID searches the space of revisions differently depending on its current motivations.

The correct revision is in fact the add-condition revision shown in Figure 8.6. We can see that `precondition-2` corresponds to the missing condition, `heat-connection ?source ?destination ?path`, for the process of Figure 8.3.

### 8.6.2 Static Selection

In a similar scenario but with the heat-connection condition satisfied (not shown), MID generates the same prediction, but again no observations are observed. If we add the background knowledge contained in the BKRB of Figure 8.7, then MID can apply another revision operator, Specialize-Pcondition. In this case, MID constructs two extra possible revisions shown in Figure 8.8, one with `aligned` being specialized to `flow-aligned`, and another with `aligned` being specialized to `heat-flow-aligned`. Both of these are

```
PROCESS:        heat-flow
Variables:      ?source
                ?destination
                ?path
Individuals:    object ?source
                object ?destination
                heat-path ?path
Pconditions:    heat-flow-aligned ?path
Qconditions:    greater-than (a (temperature ?source))
                             (a (temperature ?destination))
Relations:      Q+ (heat-flow-rate) (temperature ?source)
                Q- (heat-flow-rate) (temperature ?destination)
Influences:     I+ (heat ?destination) (a (heat-flow-rate))
Revision Log:   1 specialize preconditions
```

```
PROCESS:        heat-flow
Variables:      ?source
                ?destination
                ?path
Individuals:    object ?source
                object ?destination
                heat-path ?path
Pconditions:    flow-aligned ?path
Qconditions:    greater-than (a (temperature ?source))
                (a (temperature ?destination))
Relations:      Q+ (heat-flow-rate) (temperature ?source)
                Q- (heat-flow-rate) (temperature ?destination)
Influences:     I+ (heat ?destination) (a (heat-flow-rate))
Revision Log:   1 specialize preconditions
```

Figure 8.8: The revisions generated by MID using the BKRB above

preferred to the others by dynamic selection. Since both revisions are the result of the same single operator application, however, MID uses the static selection mechanism to distinguish between them. According to the criteria and weighting described above, MID uses conservatism, simplicity, generality, modesty and evidential support to perform the selection.

Details of the scoring in static selection are given in Table 8.5. Remember that the difference between the counts that are generated for each revision are compared against the limit values for the set of revisions being compared, and the difference between particular revisions and the limiting revision for each criterion is used instead of the original count. Here, the revisions differ on counts of disjunctions of predicates. Simplicity and modesty both prefer a smaller count, but generality prefers a larger one. Thus for generality, the limiting number is the larger one, which is for the `heat-flow-aligned` revision, while for modesty and simplicity, the limiting number is the smaller one, which is for the

| | | | | flow-aligned | | | heat-flow-aligned | | |
|---|---|---|---|---|---|---|---|---|---|
| criterion, $c$ | weight, $\omega_c$ | | difference in count, $n_c$ | $n_c$ | $n_c \times \omega_c$ | | $n_c$ | $n_c \times \omega_c$ | |
| | K | A | | | K | A | | K | A |
| conservatism | 0.8 | 0.4 | no. of changes | 0 | 0 | 0 | 1 | 0.8 | 0.4 |
| simplicity | 0.8 | 0.2 | no. of processes | 0 | | | 0 | | |
| | | | no. of disjunctions | 1 | | | 0 | | |
| | | | no. of conditions | 0 | 0.8 | 0.2 | 0 | 0 | 0 |
| generality | 0.8 | 0.1 | no. of conditions | 0 | | | 0 | | |
| | | | no. of disjunctions | 0 | | | 1 | | |
| | | | no. of effects | 0 | 0 | 0 | 0 | 0.8 | 0.1 |
| modesty | 0.2 | 0.9 | no. of conditions | 0 | | | 0 | | |
| | | | no. of disjunctions | 1 | | | 0 | | |
| | | | no. of effects | 0 | 0.2 | 0.9 | 0 | 0 | 0 |
| support | 0.1 | 0.1 | time since last revision | 0 | 0 | 0 | 0 | 0 | 0 |
| $f(\Delta)$ | | | | | 1.0 | 1.1 | | 1.6 | 0.5 |

Table 8.5: Scores for revisions in static selection.

`flow-aligned` revision.

Under knowledge motivations, Specialize-Condition: `flow-aligned` gives

$$(0.8 \times 0) + (0.8 \times (0 + 1 + 0)) + (0.8 \times (0 + 0 + 0)) + (0.2 \times (0 + 1 + 0)) + (0.1 \times 0) = 1.0$$

in comparison to Specialize-Condition: `heat-flow-aligned`

$$(0.8 \times 1) + (0.8 \times (0 + 0 + 0)) + (0.8 \times (0 + 1 + 0)) + (0.2 \times (0 + 0 + 0)) + (0.1 \times 0) = 1.6$$

Remember that only the difference between the revisions and the minimum count values for any revision is used. `flow-aligned` scores better for conservatism since it involves only one link by specialization rather than two, and for generality since it has more disjunctions of conditions and hence greater coverage) This means that it scores worse for simplicity and modesty, however. It thus scores 1.0 while `heat-flow-aligned` scores 1.6. The combined effects of conservatism and generality outweigh the effects of simplicity and modesty, and MID prefers the lower value which is associated with the specialization to `flow-aligned`. This solution will subsequently need further revision to specialize the `flow-aligned` predicate to `heat-flow-aligned`, but it has attempted to maximize the utility of the theory in maintaining a wide coverage.

Under action motivations, Specialize-Condition: `flow-aligned` gives

$$(0.4 \times 0) + (0.2 \times (0 + 1 + 0)) + (0.1 \times (0 + 0 + 0)) + (0.9 \times (0 + 1 + 0)) + (0.1 \times 0) = 1.1$$

in comparison to the minimum values of Specialize-Condition: `heat-flow-aligned`

$$(0.4 \times 1) + (0.2 \times (0 + 0 + 0)) + (0.1 \times (0 + 1 + 0)) + (0.9 \times (0 + 0 + 0)) + (0.1 \times 0) = 0.5$$

Here, `flow-aligned` scores 1.1 while `heat-flow-aligned` scores 0.5. In this case where modesty is valued more highly than generality, the `heat-flow-aligned` revision is preferred. Note that evidential support is the same for both cases, since they revise the same process. Also, the assumption has been made that consistency is not an issue here, and it has not been considered. However, we can see that the `flow-aligned` revision would be inconsistent with an appropriate history, and would require further search (or revision).

A further extended example is given in Appendix A, giving comparisons for selection with and without use of the observation *grouping* heuristic.

## 8.7 Discussion

### 8.7.1 Related Work

Of the many systems concerned with discovery, only a limited consideration has been given to selection. This section reviews a number of systems in discovery and other areas that have been implemented and which use selection criteria in some form or other.

#### Thagard's PI and ECHO

Thagard [115] identified the importance of simplicity, consilience and analogy in theory evaluation and selection, and went on to develop computational models. In PI, a program that attempts to provide a 'model of problem solving and inductive inference' [41, 117], he provided computational measures of simplicity and consilience but was unable to incorporate analogy in such a way (though still noting its significance). More recently, he has developed ECHO, a connectionist program that comparatively evaluates theories principally on the basis of a Theory of Explanatory Coherence which includes these three criteria [118, 121]. Thagard's model requires that a set of facts F be explained by a theory T and a set of auxiliary hypotheses A. The auxiliary hypotheses are claimed to be the source of complexity in that they are not part of the original theory, but are introduced as assumptions in order to explain some of the facts.

**Simplicity** In PI, one theory is simpler than another if it has a lower ratio of cohypotheses to facts explained. This ensures that a hypothesis that does not explain anything is not preferred to one that does but uses auxiliary hypotheses to do so.

ECHO's treatment is a little different. The Theory of Explanatory Coherence (TEC) states that the coherence of a theory decreases as the number of propositions increases. The propositions here are the cohypotheses that are used to explain the facts, including ad hoc assumptions. Although this mirrors the use of simplicity in PI, the connectionist nature of ECHO avoids the need for explicit values and formulae, using instead inhibitory and excitatory links to adjust coherence levels appropriately.

**Consilience** Consilience is a form of generality. A theory which explains all the facts is maximally consilient, but if this is achieved by means of auxiliary hypotheses, then it is unsatisfactory because it adds to complexity. Here we see the need for simplicity and consilience (or generality) to be taken in tandem when evaluating theories, with a decrease in simplicity — amongst other criteria — being used to offset the increase in consilience. In PI, the facts are weighted with associated importance. Thus PI considers consilience to be the sum of all the weights of the facts that it explains. If all facts are equally important, then degree of consilience can be taken to be just the number of facts explained. TEC states that what explains coheres with what is explained, and hence the more that is explained, the greater the coherence. In addition, TEC undermines the acceptability of hypotheses that explain only a small part of the relevant data. 'If many results of relevant experimental observations are unexplained, then the acceptability of a proposition $P$ that explains only a few of them is reduced.'

**Analogy** The merits of analogy as a selection criterion are particularly contentious. It is clear that analogy is significant in the formation of theories, but Thagard contends that analogy is also important in the support of hypotheses already discovered. Analogy is not used in PI, but in ECHO an analogy between two propositions will increase their coherence.

**Conservatism** Almost as a side-effect of the implementation of ECHO as a connectionist system, conservatism is tagged on to the three criteria explicitly specified above. In justification of this, Thagard suggests that conservatism is a consequence of explanatory coherence, not a separate factor. ECHO does not treat new evidence that does not cohere with existing accepted evidence as equally. For example, if a hypothesis $H_1$ explains $E_1$ and $E_2$, and subsequently a new hypothesis $H_2$ that

contradicts $H_1$ but explains $E_1$ and $E_2$ is advanced, then the system will not resettle into a state in which $H_1$ and $H_2$ receive equal activation, rather one in which $H_1$ has a higher activation than $H_2$.

From the descriptions above, it should be clear that Thagard has considered the use of selection criteria in some depth. His system is certainly very attractive, and has been demonstrated in a number of domains including legal reasoning [118], adversarial problem solving [120] [122], and a variety of scientific fields [125] [126], but is also very limited. It is only used in the evaluation of competing theories (revisions) which are presented to it by some unknown means, and ignores other issues of theory formation and revision. The connectionist implementation is immediately attractive, but leads to problems over the justification of the weights given to evidence and links in the network. They must represent something, but what and why? Part of the problem would seem to lie in the nature of the system, that it is designed only to evaluate theories in isolation from the rest of a wider process involving theory formation, revision and so on, which might yield some justification for setting weights in any particular way. Another important issue is that the evaluation that takes place in ECHO is not variable, but is designed with some vague notion of trying to acquire knowledge. As O'Rorke [83] points out, this evaluation should not be fixed, since in any real world problem, an agent's goals and priorities play important roles in evaluation.

### KEKADA

KEKADA uses a variety of ways in which to select hypotheses. First, it selects hypotheses on the basis of an associated confidence measure which, as we noted in the discussion on evaluation in Chapter 6, is simply a count of the successful and unsuccessful instances. Second, it uses a predefined ordering on hypotheses that is claimed to be on the basis of 'experience'. Third, the user can select the hypothesis. Only the confidence measure method provides reasonable automated control, and it is analogous to the evidential support criterion used here. Selection is used only to choose between hypotheses, not to constrain their generation. Clearly, the selection mechanism is very weak in KEKADA, and lacks any variation in motivation.

### COAST

COAST uses selection criteria to evaluate candidate revisions once they have been generated. Each criterion has associated with it a weight which allows the relative importance of the criteria to be varied. Rajamoney is rather vague about the use of this, and chooses default values for the context of scientific discovery arbitrarily without any justification. No discussion of other issues is provided. The criteria used are:

**Simplicity of Explanations** This criterion is based on the principle of Occam's Razor. It prefers revisions that generate simpler explanations. In this case, simplicity is defined as the number of *links* in an explanation for a given set of observations.

**Structural Simplicity (of Domain Theory)** This is based on the principle of parsimony. It prefers revisions that are simpler in terms of the number of components in a process, and the number of processes in the domain theory revision.

**Predictive Power** This is similar to the consilience of Whewell and Thagard in that it is a measure of how much a theory can account for. It prefers revisions that make more predictions, and has the merit of increasing refutability.

These criteria correspond to the simplicity and generality criteria in MID, but ignore other dimensions of selection. Also, they are limited to the second stage in the selection procedure, and play no role in constraining the revision space. Although the criteria can be weighted, it is not clear how or why they should be.

### STAHLp

STAHLp [100] was based on the STAHL program for reasoning about chemical reactions and inferring chemical models. In the event of an inconsistency in the premises of a model, the system considers revisions to them that would bring the database closer to consistency. In STAHLp, the sole selection criterion used was that of conservatism. This was superseded by REVOLVER [101] which extended STAHLp and used additional criteria for selecting between alternative models. Again there is only a minimal attempt at a serious justification of their use.

**Minimum Mutilation** This is the conservatism principle, and it prefers revisions that affect the least number of currently held beliefs in the form of chemical models.

**Complexity** Complexity has associated with it a weight that allows the preference of revisions with either a fewer or greater number of substances in the reaction. Fewer substances embodies the principle of parsimony, while more substances implies that the error was merely omitting something in observation.

**Minimum Revision** Minimum revision prefers revising premises that have been revised less often. Its main concern is to avoid cycling in the revision process.

**Same-type Assumption** This prefers revisions in which the type of substances in the premises are the same with the number of instances being different. Its purpose is to constrain the space of revisions.

These correspond to conservatism, simplicity, and criteria for efficiency and control purposes. Again, they apply only to the the second stage of selection. Both Rose and Rajamoney use selection criteria as context-dependent heuristics which, while certainly providing a useful means for evaluating proposed solutions in their particular systems, have little grounding in underlying principles.

Other discovery systems do not address theory selection explicitly. Some research on induction and machine learning has considered generality and simplicity in a limited way in concept learning (eg. [45]), but not with the breadth of dimensions considered here. Moreover, there has been little, if any at all, attention paid to the role of motivations in guiding the selection procedure.

### Story Comprehension

In discovery systems, as we have seen, there has been little, if any, consideration given to the motivations of the system in addressing the problem at hand. Some story-comprehension problems, however, particularly in explanation construction through abduction, have shown the importance of motivations in that different explanations are plausible to a greater or lesser extent depending on the goals and motivations of the agents in the stories (eg. Ram's AQUA program [96],[95], [97] and Leake's ACCEPTER program [64], [65], [66], [62]). In order to understand a story, systems must be aware of the motivations of the agents in that story, and must evaluate or select explanations of behaviour that are in accordance with them. However, the analysis of motivations is of a different kind to that proposed here. The motivations of the system are fixed and directed at explaining the behaviour of agents in the story. The motivations that are

addressed are those of the agents about whom the system is trying to construct an explanation from the events in the story, and from its own prior knowledge. Explanations are constructed not *under* motivations, but in a sense, *about* motivations, a very significant difference. Furthermore, in systems such as AQUA which have addressed motivations in this way, and for which many heuristics determining what the system should focus on exist, the problem of combining these heuristics is not addressed. Although Ram recognises this problem [97], he seems to be unaware of the implications of ignoring it. AQUA suffers in this way from overdeterminism and pursues all the alternatives, regardless of the priorities and motivations that may exist, leading to the possibility of intractability. Thus, although there is a recognition of the relevance of motivations and their use in evaluating explanations and hypotheses, the work is superficial and suffers from problems of overdeterminism and intractability. As such, it has little impact on many of the issues addressed in the reasoning and selection task that is considered here.

### 8.7.2 Conclusions

The problem of theory selection which is an important and significant one in induction and discovery, has increasingly been drawing more attention. Work on theory selection to date has focussed on the very limited task of choosing a theory from amongst many that are presented to the system, with no consideration being given to the generation of those theories. While such work is indeed relevant, and while some success has been demonstrated, the artificial limitations imposed on the problem have obscured many related issues. How are the theories to be generated in the first instance? Unconstrained hypothesis generation can lead to intractability. How may the use of selection criteria be tailored to the particular motivations of the reasoner? Different motivations may lead to different requirements on theories.

In this chapter we have considered theory selection as part of a larger model of discovery and induction. We have argued that selection is necessary in both the context of justification and the context of discovery, blurring the traditional distinction between the two. We have argued that different kinds of theories, though heavily constrained by the requirements of the inductive programme, may be required by reasoners with differing motivations. An implementation of these proposals in the MID system was described, giving specifications of a set of selection criteria and control strategies for revision and selection under different motivations.

By splitting selection into two parts, dynamic selection which constrains the process of revision, and static selection which evaluates candidate revisions once they have been generated, MID allows reasoning even when resources are limited. Only as many revisions as are necessary are generated at any point (although this can amount to an exhaustive search in the worst case). This is important because it allows extension to very large problems where the number of potential revisions is huge, without invalidating the mechanism used. Furthermore, static selection allows the introduction of new theories from external sources by alternative reasoning methods.

Necessarily, limitations exist and some further work would be profitable. In particular, the values specified for selection criteria were chosen to achieve the desired results in terms of accuracy and relevance to the problem at hand. The justification for these values was functional, but further work investigating these is warranted. It would be valuable to see if values for such criteria could be determined for human subjects, and how they might compare with those used here. Consistency checking was limited either to a complete check on the entire episodic history, or to no check at all. Further work could investigate how different degrees of consistency checking might be implemented, and the significance of the benefit that it might bring. This in turn relates to the question of how the history of events should be maintained so that performance may be optimized.

Although the analysis of motivations was necessarily crude, it was sufficient to show that motivations are both relevant and significant in the control of the selection stage. Moreover, the use of motivations provides a mechanism for manipulating selection (and other stages in the framework) so that reasoning may be both general (for knowledge) and directed (at tasks).

# Chapter 9

# Conclusions

> Science is a procedure for testing and rejecting hypotheses, not a compendium of certain knowledge.
>
> — Stephen Jay Gould, *Natural History*, Adam's Navel

## 9.1 Introduction

There are three different levels to the work described in this thesis. First, an encompassing framework for inductive discovery has been constructed within which a broad range of inductive reasoning systems can be unified. Second, a general model has been developed within this framework that describes principles and methods for a motivated inductive reasoning system. Finally, an instantiation of that model has been implemented as the MID system which reasons in simple physical domains, both for knowledge and for action.

In assessing the relevance and contribution of this work, we need to be aware of where it stands in relation to previous work. The most closely related work is Rajamoney's work on COAST. MID and COAST both use the QPT representation formalism and are therefore relatively similar in the kind of prediction and revision mechanisms used. However, MID differs from and improves on COAST in several significant ways. First, it is based on the six-stage framework so that there is a strong conceptual basis for the work. This allows the identification of those aspects of discovery not addressed in the implementation, and the organisation of the work to facilitate easy integration of those aspects at such a time when they are addressed. Second, this thesis explicitly identifies the stages of observation and evaluation which are ignored entirely by COAST, but which are addressed and implemented in this work. Third, selection in COAST simply

takes the form of three *ad hoc* fixed syntactic criteria for judging candidate theories, and does not constrain the generation of those theories. By contrast, MID justifies a larger set of selection criteria and provides a mechanism not only for applying them in both the generation and assessment of theories, but also for automatically modifying its preferences to suit the circumstances. Finally, MID provides the control mechanism of motivations to direct the reasoning appropriately, while COAST's reasoning behaviour is fixed.

Existing systems perform limited tasks in limited domains, but provide a demonstration of the potential of the rapidly expanding research program. Additional features continue to be added to discovery systems, and they continue to be extended to new domains. The limitations that exist are gradually being eliminated, as research progresses in addressing them. This thesis can be viewed in such a way — addressing existing limitations, and extending the potential of machine discovery and induction in a number of directions. In this chapter we discuss how this thesis may be evaluated, and then consider the contribution that the thesis makes to AI. Next we discuss the limitations of the work, and finally consider possibilities for further research.

## 9.2 Evaluation of MID

There is a clear need to evaluate the contribution of this thesis with regard to the field of AI and discovery. In attempting to undertake such an assessment, it is important that the criteria used are not chosen arbitrarily to highlight some features rather than others, and also that the work is evaluated in a context wider than that of any immediate application task domain. Machine discovery in particular has suffered from an excessive degree of specificity in the development of systems which replicate well-documented historical episodes in science. While such work is relevant and necessary, it is recognised that it must be complemented by more generally applicable work [111, 103]. Here, we consider the work described in this thesis in this regard.

We shall begin this discussion by considering the generality (and applicability) of the techniques and methods described here. There are a number of points where a discussion of generality is appropriate: the problem, the representation, the inference mechanism, the search and heuristics. First, the problems that are addressed by this research are not isolated examples from the history of science that capture a very specific type of solution,

but are more general problems with knowledge in a variety of contexts limited primarily by the manner of its explicit representation for use by computer systems. Thus our concern is with knowledge-based systems in the very broadest sense. Although this does not ground the work in a well-documented historical episode, the broad context of computer knowledge-based systems provides a more general basis, one which is immediately relevant, but which is no less well defined.

Second, the representation of the knowledge that is the subject of the investigations here, both in acquisition and revision, sits comfortably in a much larger field that is concerned with issues of representation *per se*. Qualitative Process Theory (QPT) is a well-defined formalism on which research is still actively being undertaken, and which provides a language suitable for use in qualitative reasoning about physical domains. The avoidance of *ad hoc* representations is not coincidental, but a conscious effort to contribute to the generality and applicability of this research as a whole. Nevertheless, translations to other formalisms (such as simple production rules, for example,) are readily made by discussing the QPT formalism at different levels. The representation formalism was discussed at a level of detail necessary for a complete appreciation of the model and its implementation, but abstractions to more general terms (such as conditions and effects) were used whenever possible and appropriate. Such abstractions allow the methods and main algorithms used here to be modified to cover different formalisms.

Third, the inference mechanisms used here are relatively straightforward, and are easily applied to other representations, domains and contexts. Ranging from the simple pattern matcher used for generating predictions through the methods for evaluating evidence to those for revising and selecting theories, the reasoning has been kept as general as possible, and when this has not been possible, the discussion in this thesis provides a basis for extension and application to other instances. This has been achieved by including two levels of description and analysis: one at a general level in terms of underlying principles regardless of instantiations in a particular implementation, and the other at the specific level of the implementation instantiation itself, showing the path from the general to the specific in a particular case.

In line with this, the search control through the use of heuristic selection criteria was developed both in terms of the general underlying principles, and in terms of the specifics necessary for an implementation based on QPT. Thus although some of the particular heuristics implemented are based around the structure of the QPT formalism,

abstraction to higher levels both in an initial discussion and in a subsequent consideration of the implemented criteria provides the information necessary for other instantiations of the same criteria in different circumstances.

While efforts have been made to develop general methods, this has not always been uniformly possible, and the generality that has been possible is unavoidably limited in some ways. What can be claimed, however, is that there is an awareness of the limitations of the work described here, and these limitations have been stated (as far as possible). Moreover, in developing this work, the use of abstraction provides at least a guide to the way in which these limitations may be overcome in contexts that have not been considered here.

## 9.3 Contribution to AI

### 9.3.1 The Six-Stage Framework

In providing a sound conceptual base from which to work, a framework for induction and discovery was developed. The framework, though based on the scientific discovery paradigm, identifies the principal stages involved in inductive reasoning of various kinds, and not just scientific discovery. As AI has developed, a fragmentation has occurred and barriers have been created between different subfields. An important aim in constructing the framework was to surmount these barriers, and to allow different models and paradigms of reasoning to be unified under a single encompassing structure.

In addition, the framework enables the construction of a model and implementation that is grounded in a broader and more complete conceptual structure. Six distinct stages of the framework were identified: prediction, experimentation, observation, evaluation, revision and selection. The model described in this dissertation concentrated on the internal stages of the framework, but by being grounded in the conceptual structure of the framework, avoided the common pitfall of isolation from those elements not addressed in detail. The model was developed within the six-stage framework with an appreciation and understanding of the relevance of those other elements.

The six-stage framework provides a clear viewpoint from which to consider other paradigms, models and implementations. It is complementary to the framework of Cheng [6] which proposes a hierarchical structure relating paradigms, setups, and tests in experimentation, and hypotheses, models and instances in theory. In this hierarchy, the

six-stage framework assumes the most abstract level of paradigm or hypothesis. Models developed within that framework thus become setups or models, and implementations are tests or instances. In this respect, all of the work described here fits neatly into Cheng's hierarchy.

### 9.3.2 Motivations

Many critics of AI argue that the real power of the technology lies in the hands of the programmers who manipulate and tweak their designs to suit circumstances. They argue that extensive programmer (or user) intervention invalidates much of the benefit to be gained from this work by demanding modifications for each novel situation. One of the ways in which this can be countered is by making intelligent systems autonomous, by giving them the power to control themselves. Furthermore, it has been suggested that the difference between learning and discovery is autonomy. A significant contribution of this thesis has been to show that a degree of autonomy can be achieved through the modelling and use of the motivations of a reasoning agent.

A stream of recent research in AI has focussed on the modelling of autonomous agents in distributed problem solving through cooperation, negotiation, and so on. The aim of the current research has a different focus that is not centred on the modelling of autonomous agents or their qualities, but on *using* aspects of agents to provide control strategies for inductive reasoning. Consequently, in modelling motivations in MID, a limited representation was developed, with work directed at showing how these motivations affect reasoning and action in the world, rather than how action and reasoning in the world affect motivations. Moreover, the reasoning and action addressed were not the simple behavioural-response kind, but of higher level reasoning strategies. In this way, the current research complements the related but distinct work on modelling artificial agents and creatures.

Despite the simplicity of the model of motivations developed, this work has shown that even such a simple model provides an effective means for controlling the different elements in inductive discovery, and for allowing different kinds of reasoning to be derived from a single set of components. Clearly, more work remains to be done on motivations, but we see in MID a demonstration of the potential gain in flexibility and expressiveness that can result from their explicit representation and use.

### 9.3.3 The Model

Work on developing a model and implementation concentrated on the internal stages of the framework. In particular, the stages of evaluation, revision and selection were considered in detail.

#### Evaluation

Evaluation is an important element in all kinds of systems in which an acknowledgement is made of the possibility of error or failure. Traditionally, it has received little attention, assuming perfect evidence and fixed standards of judgement. In inductive reasoning, evaluation takes on even greater significance because the evidence is the sole (or primary) fuel to the reasoning engine. This work makes two main contributions that can be identified: the recognition and evaluation of different dimensions of uncertainty; and the variable standards of acceptability according to the needs of the reasoning agent.

Considerations of error and uncertainty have been lacking in very many systems which rely on evidence from an imperfect external world. When they have been addressed, however, the consideration has usually been restricted to simple tolerance limits for the numerical *accuracy* of the evidence. This is inadequate for systems which aim to function in any world that admits a greater variety of uncertainty. At a basic level, uncertainty can be split into that which results from the phenomenon itself, and that which results from observation of the phenomenon. Often, an external source is needed for observation, both in science and in ordinary life. In computing, for example, the issues involved in evaluation are becoming increasingly important as the use of distributed systems of all sorts grows rapidly [78].

The acceptance of evidence must also be made on the basis of the need for that evidence. In situations where the potential consequences of an incorrect inference (due to faulty evidence) are serious, high quality evidence is demanded. In situations that have little significance, poor quality evidence may be adequate. Through modelling motivations, an indication of the importance of the situation to the reasoning agent can be derived, allowing the requirements placed on evidence to be tailored to the relevant needs. This work proposes a model and implementation for evaluation incorporating these concerns, and shows how motivations provide the necessary guidance.

### Revision

Much research is currently being undertaken into theory revision. This work has emphasised the need for clear specification of revision operators, and of revision strategies. In particular, it is important that all revisions are considered, and that all combinations of revision operators are considered in generating revisions. Any implicit constraints that are imposed on the generation of revisions must be clearly stated. We have described the requirements of theory revision systems and justified the different revisions that are possible, explaining why and in what circumstances the various revisions are appropriate. In particular, we have been careful not to rule out revisions in a prejudiced way, but to evaluate each according to its merit. In order to avoid problems of combinatorial explosion, however, revision is constrained by selection.

### Selection

Traditionally, theory selection has involved the generation of multiple revisions, and then their evaluation according to some fixed metric so that the best or most plausible theory may be identified. This work departs from that view of selection in two significant ways.

First, the generation of theories or revisions must be constrained so that the process is manageable. If all possible revisions are actually to be generated, then the problem becomes intractable, particularly in non-trivial domains where the number of potential revisions can be excessive. Thus much existing work does not extend beyond the immediate experimental domain. In MID, selection is of two kinds: dynamic selection which constrains the revision procedure; and static selection which evaluates candidate revisions once they have been generated. Dynamic selection avoids the need to generate all revisions since the revision operators themselves are ordered so that the best or most plausible revisions are generated first. (This can, however, amount to an exhaustive search in a worst case scenario.) It is important because it allows extension to richer domains and larger theories without invalidating the mechanism used. Static selection is used when multiple revisions are generated by a single operator or when multiple operators are used. This is also important if the system is not to be closed to theories generated by external sources using alternative reasoning methods such as analogy. Static selection criteria provide a means of evaluating such candidate revisions relative to each other so that new theories may be introduced.

Second, the relative importance of the selection criteria used can be varied to suit the purpose of the reasoner. Theories are valued differently with different motivations, and the kind of revision preferred will depend on this. If there is an immediate need for a theory so that action can be taken, it may not be possible to find the best theory given the time constraints. The requirements of such a theory will be different to those of a theory that is not subject to such constraints. Such considerations are an important part of selection. In contrast to existing systems, we explicitly address these issues through the use of motivations to determine value sets for selection criteria.

### 9.3.4 Resource Bounds

An important claim of this work is that scientific reasoning should apply just as well to common everyday situations as to the pursuit of science. Inductive discovery involves reasoning that is suitable in very many domains and contexts. Clearly, there are differences between the special case of science and other less controlled environments, but these are increasingly being addressed. Researchers in scientific discovery are turning their attention to the problem of knowledge discovery in databases, for example, in which the issues are related but not identical (eg. [24], [136], [74], [139]). In scientific discovery, the data is typically dense and of high quality, obtained through directed experimentation in a well controlled environment. In other domains, however, data is rarely as good, and the reasoning must consequently be biased differently.

In addition, the performance of a reasoning agent in science contrasts strongly with the responsiveness demanded of reasoning agents in other domains which may need to react quickly. The lengthy deliberation that is required to reach the best solution may be beyond the ability of the agent to perform in a limited time and with limited resources. The work described here addresses these issues by using motivations to control the amount of inference that can be undertaken. When resources are limited, 'quick and dirty' solutions may be best, but when they are unconstrained, the optimal reasoning strategy can be pursued. Thus MID provides a means for adapting the reasoning so that it is appropriate for the relevant domain, and for the abilities of the reasoning agent in that domain.

## 9.4 Limitations

Inevitably, there are a number of distinct limitations to the work that is described here. At the broadest level, and within the context of the six-stage framework, the focus of the research undertaken has been on the internal stages of the framework. Though external issues have been noted, only the internal stages have been considered in detail. Primarily, this relates to the stage of experimentation which has not been modelled here at all, and to observation which is considered in a very limited way.

The emphasis on internal factors is also reflected in the modelling of motivations. Our concern with motivations is limited to their effect on reasoning, and not with the modelling of artificial agents. Thus this work does not address issues of how motivations vary in response to their environment or to changes in that environment. These limitations impose a stronger focus because the boundaries are clearly defined and recognised.

In addition, it should be noted that the implementation of the MID system is not intended to be a fully functional system capable of actual inductive discovery in a real world, rather a demonstration that the ideas and mechanisms proposed are computationally possible and effective. Thus MID is a prototype, limited by design and intention rather than limited by inadequacy. With further development of the system, however, many of these limitations would be removed.

More detailed discussions of the limitations of the different elements of this work can be found in the individual chapters.

## 9.5 Future Work

This research has opened up many avenues for further research, both in the development of the ideas presented, and in their more complete integration with related aspects. We identify several major directions that such research may pursue.

First, we note the boundaries of the current work already discussed, and the restriction to the internal stages of the framework. Further work could usefully pursue research into experimentation strategies, and in integrating existing strategies and systems with the current work. An interesting problem would be to see how well the methods developed here for reasoning about theoretical knowledge could be used to reason about experimental knowledge used in the directed design and construction of experiments.

In addition, a much stronger representation of the external environment could be

developed, together with a more detailed model of motivations specifying the way in which motivations respond to it. Alternatively, the current work could be integrated with existing mechanisms for determining motivational values such as those developed by Maes [77]. This would allow motivational strength values to be determined directly rather than by interface with the programmer.

As it exists currently, MID is designed to acquire knowledge and eliminate errors from its domain theory without preference or bias. It is interested in everything that satisfies these goals. Further work could extend MID by introducing additional goals to be satisfied. This would have implications for a number of elements.

The introduction of these extra goals may constrain prediction, for example, making certain predictions more relevant than others. In such a case, effort would need to be directed at predicting (and hence reasoning about) those anomalies that have greater utility, say, than other anomalies.

In selection, too, the introduction of additional goals would be significant. The selection criteria used are general criteria that are intended for use solely in the acquisition of knowledge. In order to satisfy other new goals, the criteria would need to be extended to include more domain-specific and context-specific criteria. The story understanding programs of Leake and Ram discussed in Chapter 8 provide an indication of the direction in which MID would need to be extended.

The existing selection criteria cannot be claimed to be a definitive set of general criteria, but they are sufficient for the goals of the MID system. Further investigation is warranted, however, in the values given to the criteria when reasoning under knowledge and action motivations. Although the values given here are justified by pragmatic concerns, it would be beneficial to see if there is a correlation with these value sets in human reasoning, and whether alternative value sets are possible.

## 9.6 Conclusion

In this thesis, a new dimension has been added to computational discovery systems through the use of motivations as a control strategy, providing evidence of the potential of such systems, and of their applicability and richness. Despite the simplicity of the representational formalism, significant control is exercised, and different modes of inductive reasoning can be derived. Thus a single set of processes is shown to be sufficient

to produce a variety of reasoning behaviours as appropriate and necessary according to motivations and circumstances.

Though situated firmly in the discovery camp, this work is also relevant to other areas of AI. In considering the ability to deal with bounds on resources we note, for example, that some of the requirements suggested by Waltz [131] for intelligent robots are satisfied by the MID system. These include the use of innate drive and evaluation systems to provide moment-to-moment guidance for action, and the adoption of the best existing alternative without further evaluation in emergencies. Both of these are significant aspects of this work, which can be applied to many other areas.

Scientific reasoning provides a basis for predicting and controlling our environment. Inductive discovery is a mode of scientific reasoning which has achieved much success in the history of science, and which has attracted efforts by AI researchers to emulate that success. In this pursuit, computer programs have been developed that attempt to replicate the *reported* reasoning that has led to significant scientific discoveries. The concern of this thesis has not been to simulate a particular historical episode, but to show how inductive discovery can be applied to real-world problems that arise both in scientific and non-scientific domains. At a meta level, this research can be regarded as a theory that has survived our attempts to refute it through experimentation thus far. Ideally, however, we would want to develop a recursive structure with systems reasoning about themselves, taking account of success and failure, and playing a part in the never-ending struggle for knowledge in the best tradition of scientific progress.

# Appendix A

# An Extended Example

The example of MID's operation given here is based on the data used by Rajamoney [91] for his osmosis example. First we give the complete domain theory used.

## A.1 Domain Theory

```
Process Name:   Solution
Individuals:    contained-liquid ?solution
Pconditions:    soluble (solute-of ?solution) (solvent-of ?solution)
Qconditions:    greater-than (a (amount-of (solute-of ?solution))) 0
Relations:      Q+ (concentration ?solution) (amount-of (solute-of ?solution))
                Q- (concentration ?solution) (amount-of (solvent-of ?solution))
                Q+ (amount-of ?solution) (amount-of (solvent-of ?solution))
Influences:

Process Name:   Evaporation
Individuals:    contained-liquid ?liquid
                contained-gas ?vapor
Pconditions:    connection ?liquid ?vapor
                open-container container ?liquid
Qconditions:
Relations:      Q+ (evaporate-rate ?evaporation-rate)
                                              (contact-area ?liquid ?vapor)
Influences:     I- (amount-of ?liquid) (a (evaporation-rate))
                I+ (amount-of ?vapor) (a (evaporation-rate))

Process Name:   Condensation
Individuals:    contained-gas ?vapor
                contained-liquid ?liquid
Pconditions:    connection ?liquid ?vapor
                open-container container ?liquid
Qconditions:
Relations:      Q + (condensation-rate) (contact-area ?liquid ?vapor)
Influences:     I+ (amount-of ?liquid) (a (condensation-rate))
                I- (amount-of ?vapor) (a (condensation-rate))
```

```
Process Name:   Absorption
Individuals:    solid ?solid
                contained-liquid ?liquid
Pconditions:    connection ?solid ?liquid
                absorbent ?solid
Qconditions:    less-than (a (absorbed-liquid-of ?solid))
                          (a (maximum-absorbed-liquid-of-point ?solid))
Relations:      Q+ (absorption-rate) (contact-area ?solid ?liquid)
Influences:     I- (amount-of ?liquid) (a (absorption-rate))

Process Name:   Release
Individuals:    solid ?solid
                contained-liquid ?liquid
Pconditions:    connection ?solid ?liquid
                absorbent ?solid
Qconditions:    greater-than (a (absorbed-liquid-of ?solid))
                             (a (minimum-absorbed-liquid-of-point ?solid))
Relations:      Q+ (release-rate) (contact-area ?solid ?liquid)
Influences:     I+ (amount-of ?liquid) (a (release-rate))

Process Name:   Fluid-flow
Individuals:    contained-fluid ?source
                contained-fluid ?destination
                path ?path
Pconditions:    path-connection ?source ?destination ?path
                fluid-flow-aligned ?path
Qconditions:    greater-than (a (pressure ?source))
                                        (a (pressure ?destination))
Relations:      Q+ fluid-flow-rate (pressure ?source)
                Q- (fluid-flow-rate) (pressure ?destination)
Influences:     I+ (amount-of ?destination) (a (fluid-flow-rate))
                I- (amount-of ?source) (a (fluid-flow-rate))

Process Name:   Add-solute
Individuals:    contained-solid ?solute-source
                contained-solid ?solute-destination
Pconditions:    transfer-connection ?solute-source ?solute-destination
                transferable?  ?solute-source ?solute-destination
Qconditions:
Relations:
Influences:     I- (amount-of ?solute-source) (a (add-solute-rate))
                I+ (amount-of ?solute-destination) (a (add-solute-rate))
```

## A.2 Background Knowledge Rule Base

```
flow-aligned        →  aligned
fluid-flow-aligned  →  flow-aligned
heat-flow-aligned   →  flow-aligned
```

## A.3 Successful Prediction

MID successfully predicts the effects of the following scenarios. These provide MID with a limited though important history so that historical consistency may be checked.

```
Scenario:    fluid-flow-works-scenario
Individuals: std-fluid1
             std-fluid2
             std-path1
Facts:       contained-fluid std-fluid1
             contained-fluid std-fluid2
             path std-path1
             path-connection std-fluid1 std-fluid2 std-path1
             fluid-flow-aligned std-path1
             greater-than (a (pressure std-fluid1)) (a (pressure std-fluid2))
```

MID successfully predicts that the `fluid-flow` process is active and that the amount of `std-fluid2` increases, while the amount of `std-fluid1` decreases. A trace is not shown.

```
Scenario:    fluid-flow-fails1-scenario
Individuals: std-fluid1
             std-fluid2
             std-path1
Facts:       contained-fluid std-fluid1
             contained-fluid std-fluid2
             path std-path1
             path-connection std-fluid1 std-fluid2 std-path1
             not-fluid-flow-aligned std-path1
             greater-than (a (pressure std-fluid1)) (a (pressure std-fluid2))
```

MID makes no predictions. No changes are observed.

```
Scenario:    absorption-works-scenario
Individuals: solid1
             liquid1
Facts:       solid solid1
             contained-liquid liquid1
             connection solid1 liquid1
             absorbent solid1
             less-than (a (absorbed-liquid-of solid1))
                       (a (maximum-absorbed-liquid-of-point solid1))
```

A decrease in the amount of `liquid1` is observed.

```
Scenario:    absorption-fails-scenario
Individuals: solid1
             liquid1
Facts:       solid solid1
             contained-liquid liquid1
             connection solid1 liquid1
             less-than (a (absorbed-liquid-of solid1))
                       (a (maximum-absorbed-liquid-of-point solid1))
```

No changes are observed.

```
Scenario:     release-works-scenario
Individuals:  solid1
              liquid1
Facts:        solid solid1
              contained-liquid liquid1
              connection solid1 liquid1
              absorbent solid1
              greater-than (a (absorbed-liquid-of solid1))
                           (a (minimum-absorbed-liquid-of-point solid1))
```

An increase in the amount of liquid1 is observed.

```
Scenario:     release-fails-scenario
Individuals:  solid1
              liquid1
Facts:        solid solid1
              contained-liquid liquid1
              connection solid1 liquid1
              greater-than (a (absorbed-liquid-of solid1))
                           (a (minimum-absorbed-liquid-of-point solid1))
```

No changes are observed.

## A.4 Correcting an Anomaly

Now MID is given a much more complicated scenario. This describes a situation in which the novel process of osmosis occurs. MID does not know about osmosis and consequently does not predict any changes.

```
Scenario:     osmosis-scenario1
Individuals:  solution1
              solution2
              vapor1
              vapor2
              wall
              partition
              wall-path
              partition-path
Facts:        contained-liquid solution1
              contained-liquid solution2
              contained-gas vapor1
              contained-gas vapor2
              contained-fluid solution1
              contained-fluid solution2
              contained-fluid vapor1
              contained-fluid vapor2
              solid wall
              solid partition
              path wall-path
              path partition-path
              connection solution1 vapor1
```

```
connection solution2 vapor2
connection wall solution1
connection wall solution2
connection partition solution1
connection partition solution2
path-connection solution1 solution2 partition-path
path-connection solution1 solution2 wall-path
soluble (solute-of solution1) (solvent-of solution1)
soluble (solute-of solution2) (solvent-of solution2)
greater-than (a (amount-of:solute-of solution1)) 0
greater-than (a (amount-of:solute-of solution2)) 0
greater-than (a (concentration solution1))
             (a (concentration solution2))
greater-than (a (absorbed-liquid-of wall))
             (a (minimum-absorbed-liquid-of-point wall))
less-than (a (absorbed-liquid-of wall))
          (a (maximum-absorbed-liquid-of-point wall))
```

MID first attempts to derive predictions from the domain theory and scenario description.

```
 Matching domain processes against scenario description ...
PROCESS: solution
Variables:   solution1
Individuals: contained-liquid solution1
Pconditions: soluble (solute-of solution1) (solvent-of solution1)
Qconditions: greater-than (a (amount-of:solute-of solution1)) 0
Relations:   Q+ (concentration solution1) (amount-of:solute-of solution1)
             Q- (concentration solution1) (amount-of:solvent-of solution1)
             Q+ (amount-of solution1) (amount-of:solvent-of solution1)
Influences:
Revision Log:

PROCESS: solution
Variables:   solution2
Individuals: contained-liquid solution2
Pconditions: soluble (solute-of solution2) (solvent-of solution2)
Qconditions: greater-than (a (amount-of:solute-of solution2)) 0
Relations:   Q+ (concentration solution2) (amount-of:solute-of solution2)
             Q- (concentration solution2) (amount-of:solvent-of solution2)
             Q+ (amount-of solution2) (amount-of:solvent-of solution2)
Influences:
Revision Log:

Inactive: evaporation
Inactive: condensation
Inactive: absorption
Inactive: release
Inactive: fluid-flow
Inactive: add-solute

No predictions can be generated!
--- Active processes have no influences (effects).
```

Note that although the solution process is active and can be instantiated with both

solution1 and solution2, it has no influences (or direct effects) which can become predictions.

MID is now provided with observations of changes to quantities, resulting in an anomaly. Very different reasoning behaviours are possible here. Below we show MID's output when reasoning for action and when reasoning for knowledge, with and without the *grouping* heuristic. All possible revisions are given. Those revisions that follow the first *consistent* revision do not need to be generated, but are shown for completeness of demonstration. In subsequent traces, a condensed trace of output is given, showing only relevant details. Revisions that are generated after the first consistent revision will be condensed.

First, we show MID reasoning for knowledge without the *grouping* heuristic. Note that the motivation for knowledge is higher than the action motivation.

### A.4.1 Without Grouping Observations

```
Current Motivations:
<knowledge,0.6,Fixed>
<action,0.4,Variable>

Maximum confidence: 0.9, and minimum confidence: 0.1
Importance: 0.6
Urgency: 0.2 with a limit of: 0.7

Accuracy: 0.8                    Credibility: 0.9

Enter reliability of observer: .9
Enter trustworthiness of observer: 1

Acceptance threshold: 0.6 ----> Action point: 0.6
Confidence: 0.648

Confidence exceeds action point - observations are accepted!

Enter observations (end with return alone):
> increase (amount-of solution2)
> decrease (amount-of solution1)
>
-----
Anomalous observations: theory refuted!
```

The evidence is acceptable, but the observations do not match the predictions. The theory must be revised.

```
Current Motivations:
<knowledge,0.6,Fixed>
```

```
<action,0.4,Variable>

Searching for revisions ...

Revision #1
HISTORICALLY INCONSISTENT --- clash with: fluid-flow-fails1-scenario
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
PROCESS: fluid-flow
Variables:   ?source
             ?destination
             ?path
Individuals: contained-fluid ?source
             contained-fluid ?destination
             path ?path
Pconditions: path-connection ?source ?destination ?path
Qconditions:
Relations:   Q+ (fluid-flow-rate) (pressure ?source)
             Q- (fluid-flow-rate) (pressure ?destination)
Influences:  I+ (amount-of ?destination) (a (fluid-flow-rate))
             I- (amount-of ?source) (a (fluid-flow-rate))
Revision Log: 8 delete conditions

No change: add-solute

--------------------
Revision #2
HISTORICALLY INCONSISTENT --- clash with: absorption_fails_scenario
No change: solution
No change: evaporation
No change: condensation
PROCESS: absorption
Variables:   ?solid
             ?liquid
Individuals: solid ?solid
             contained-liquid ?liquid
Pconditions: connection ?solid ?liquid
Qconditions: less-than (a (absorbed-liquid-of ?solid))
                       (a (maximum-absorbed-liquid-of-point ?solid))
Relations:   Q+ (absorption-rate) (contact-area ?solid ?liquid)
Influences:  I- (amount-of ?liquid) (a (absorption-rate))
Revision Log: 8 delete conditions

PROCESS: release
Variables:   ?solid
             ?liquid
Individuals: solid ?solid
             contained-liquid ?liquid
Pconditions: connection ?solid ?liquid
Qconditions: greater-than (a (absorbed-liquid-of ?solid))
                          (a (minimum-absorbed-liquid-of-point ?solid))
Relations:   Q+ (release-rate) (contact-area ?solid ?liquid)
```

```
Influences:  I+ (amount-of ?liquid) (a (release-rate))
Revision Log: 8 delete conditions

No change: fluid-flow
No change: add-solute

---------------------
Revision #3
HISTORICALLY INCONSISTENT --- clash with: release-fails-scenario
No change: solution
No change: evaporation
No change: condensation
No change: absorption
PROCESS: release
Variables:   ?solid
             ?liquid
Individuals: solid ?solid
             contained-liquid ?liquid
Pconditions: connection ?solid ?liquid
Qconditions: greater-than (a (absorbed-liquid-of ?solid))
                          (a (minimum-absorbed-liquid-of-point ?solid))
Relations:   Q+ (release-rate) (contact-area ?solid ?liquid)
Influences:  I+ (amount-of ?liquid) (a (release-rate))
Revision Log: 8 delete conditions

No change: fluid-flow
No change: add-solute
PROCESS: process8
Variables:   ?var-15
Individuals: contained-liquid ?var-15
Pconditions: precondition-osmosis-scenario ?var-15
Qconditions:
Relations:
Influences:  I- (amount-of ?var-15) (a (process8-rate))
Revision Log: 8 new_process

---------------------
Revision #4
HISTORICALLY INCONSISTENT --- clash with: absorption_fails_scenario
No change: solution
No change: evaporation
No change: condensation
PROCESS: absorption
Variables:   ?solid
             ?liquid
Individuals: solid ?solid
             contained-liquid ?liquid
Pconditions: connection ?solid ?liquid
Qconditions: less-than (a (absorbed-liquid-of ?solid))
                       (a (maximum-absorbed-liquid-of-point ?solid))
Relations:   Q+ (absorption-rate) (contact-area ?solid ?liquid)
Influences:  I- (amount-of ?liquid) (a (absorption-rate))
Revision Log: 8 delete conditions

No change: release
```

```
No change: fluid-flow
No change: add-solute
PROCESS: process8
Variables:   ?var-15
Individuals: contained-liquid ?var-15
Pconditions: precondition-osmosis-scenario ?var-15
Qconditions:
Relations:
Influences:  I+ (amount-of ?var-15) (a (process8-rate))
Revision Log: 8 new_process

---------------------
Revision #5
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
No change: fluid-flow
No change: add-solute
PROCESS: process8
Variables:   ?var-15
Individuals: contained-liquid ?var-15
Pconditions: precondition-osmosis-scenario ?var-15
Qconditions:
Relations:
Influences:  I+ (amount-of ?var-15) (a (process8-rate))
Revision Log: 8 new_process

PROCESS: process9
Variables:   ?var-16
Individuals: contained-liquid ?var-16
Pconditions: precondition-osmosis-scenario ?var-16
Qconditions:
Relations:
Influences:  I- (amount-of ?var-16) (a (process9-rate))
Revision Log: 8 new_process

---------------------
Revision #6
HISTORICALLY INCONSISTENT --- clash with: fluid-flow-fails1-scenario
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
PROCESS: fluid-flow
Variables:   ?source
             ?destination
             ?path
Individuals: contained-fluid ?source
             contained-fluid ?destination
             path ?path
Pconditions: path-connection ?source ?destination ?path
Qconditions:
```

```
Relations:    Q+ (fluid-flow-rate) (pressure ?source)
              Q- (fluid-flow-rate) (pressure ?destination)
Influences:   I+ (amount-of ?destination) (a (fluid-flow-rate))
              I- (amount-of ?source) (a (fluid-flow-rate))
Revision Log: 8 delete conditions

No change: add-solute

--------------------
Revision #7
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
No change: fluid-flow
No change: add-solute
PROCESS: process8
Variables:    ?var-15
              ?var-16
Individuals: contained-liquid ?var-15
              contained-liquid ?var-16
Pconditions: precondition-osmosis-scenario ?var-15 ?var-16
Qconditions:
Relations:
Influences:   I+ (amount-of ?var-15) (a (process8-rate))
              I- (amount-of ?var-16) (a (process8-rate))
Revision Log: 8 new_process
```

With an anomalous observation failure, we attempt to revise the theory so that the observations are caused as a result. This can be done by modifying the effects of active processes to include those that match the observations, modifying the conditions of inactive processes with effects entailing the observations, or creating new processes. The only active process is `solution`, which is instantiated separately by both observed quantities. Since the direction of change for each is different, a modification of the `solution` process is not possible. The second possibility is to modify the conditions of inactive processes which could cause appropriate predictions. These include the `fluid-flow`, `release` and `absorption` processes, which demand the modification of preconditions and quantity conditions appropriately. The last possible revision is to create new processes. The selection vectors for the operators are shown in Table A.1. The value weights for revision when reasoning under knowledge motivations are:

$\mathcal{W}_{knowledge} =< 0.8, 0.8, 0.8, 0.2, 0.1, 1.0 >$

Using the first four of these for dynamic selection, MID calculates an ordering on these revision operators shown in Table A.1

The ordering on these operators determines the order of revision. Thus the first to

| *revision operator* | *selection vector* | *knowledge* | *action* |
|---|---|---|---|
| Delete-Condition | $< -1, 1, 1, -1 >$ | 0.6 | -1.0 |
| Negate-Condition | $< -1, 0, 0, 0 >$ | -0.8 | -0.4 |
| Generalize-Condition | $< 1, -1, 1, -1 >$ | 0.6 | -0.6 |
| Add-Effect | $< 1, 0, 1, -1 >$ | 1.4 | -0.4 |
| Invert Effect | $< -1, 0, 0, 0 >$ | -0.8 | -0.4 |
| Generalize Effect | $< 1, 0, 1, -1 >$ | 1.4 | -0.4 |
| New-Process | $< -1, -1, 0, 1 >$ | -1.4 | 0.3 |

Table A.1: Selection vectors for relevant revision operators.

be considered are **add-effect** and **generalize-effect**, but both of these cannot generate revisions as discussed above. Next are **delete-condition** and **generalize-condition**, but of these only **delete-condition** is applicable, and it can apply to `release` and `fluid-flow`. (It also applies to `absorption`, but observations are addressed in order, the first of which, `increase (amount-of solution2)`, is not an effect of `absorption`.) The **negate-condition** and **invert-effect** operators are both not acceptable because the inversion of relevant conditions do not appear in the scenario description, and because there are no appropriate active processes. Finally, the last operator is the **new-process** operator.

Since the **delete-condition** operator can be applied to two processes, static selection is used to distinguish between them. In the case of `release`, only one precondition needs to be deleted, and with `fluid-flow`, one precondition and one quantity condition are deleted. Thus, the `release` revision has an extra condition but one less change than the `fluid-flow` revision. In terms of the criteria, `release` scores better for conservatism and modesty, but `fluid-flow` scores better for both generality and simplicity. Remember that the difference between the counts that are generated for each revision are compared against the limit values for the set of revisions being compared, and the difference between particular revisions and the limiting revision for each criterion is used instead of the original count. Here, the revisions differ on counts of conditions. Simplicity and generality both prefer a smaller count, but modesty prefers a larger one. Thus for modesty, the limiting number is the smaller one, which is for the `release` revision, while for generality and simplicity, the limiting number is the larger one, which is for the `fluid-flow` revision. Using the weights above, we get the ordering values shown in Table A.2, where `release` scores 1.6, and `fluid-flow` scores 1.0 and is thus preferred.

| | | | | release | | | fluid-flow | | |
|---|---|---|---|---|---|---|---|---|---|
| criterion, $c$ | weight, $\omega_c$ | | difference in count, $n_c$ | $n_c$ | $n_c \times \omega_c$ | | $n_c$ | $n_c \times \omega_c$ | |
| | K | A | | | K | A | | K | A |
| conservatism | 0.8 | 0.4 | no. of changes | 0 | 0 | 0 | 1 | 0.8 | 0.4 |
| simplicity | 0.8 | 0.2 | no. of processes | 0 | | | 0 | | |
| | | | no. of disjunctions | 0 | | | 0 | | |
| | | | no. of conditions | 1 | 0.8 | 0.2 | 0 | 0 | 0 |
| generality | 0.8 | 0.1 | no. of conditions | 1 | | | 0 | | |
| | | | no. of disjunctions | 0 | | | 0 | | |
| | | | no. of effects | 0 | 0.8 | 0.1 | 0 | 0 | 0 |
| modesty | 0.2 | 0.9 | no. of conditions | 0 | | | 1 | | |
| | | | no. of disjunctions | 0 | | | 0 | | |
| | | | no. of effects | 0 | | | 0 | 0.2 | 0.9 |
| support | 0.1 | 0.1 | time since last revision | 0 | 0 | 0 | 0 | 0 | 0 |
| $f(\Delta)$ | | | | | 1.6 | 0.3 | | 1.0 | 1.3 |

Table A.2: Scores for revisions in static selection.

| *Revision* | *first step* | *second step* |
|---|---|---|
| Revision #1 | **delete-condition fluid-flow** | — |
| Revision #2 | **delete-condition release** | **delete-condition absorption** |
| Revision #3 | **delete-condition release** | **new-process** |
| Revision #4 | **new-process** | **delete-condition absorption** |
| Revision #5 | **new-process** | **new-process** |
| Revision #6 | **delete-condition fluid-flow** | — |
| Revision #7 | **new-process** | — |

Table A.3: Final ordering of revisions under knowledge motivation.

The final ordering on revisions is thus the one shown in Table A.3. Note that where only one step revision occurs, it accounts for both observations.

We can see that the first revision that is consistent with prior evidence is revision #5. Each of the preceding revisions must be checked against each of the scenarios in the history to ensure consistency, demanding significant resources. Note that revisions #3 and #6 are the same, but that revision #3 arises through attempting to accommodate the first observations, and #6 arises through attempting to accommodate both observations together. Both revisions #5 and #7 characterize osmosis, but #5 uses two distinct processes to do so.

If we need to take action immediately, the resources demanded may not be available. Below, the same scenario is given, but this time the desire for action motivates the rea-

| *Revision* | *first step* | *second step* |
|---|---|---|
| Revision #1 | **new-process** | **new-process** |
| Revision #2 | **new-process** | **delete-condition absorption** |
| Revision #3 | **delete-condition release** | **new-process** |
| Revision #4 | **delete-condition release** | **delete-condition absorption** |
| Revision #5 | **delete-condition fluid-flow** | — |
| Revision #6 | **new-process** | — |
| Revision #7 | **delete-condition fluid-flow** | — |

Table A.4: Final ordering of revisions under action motivation.

soning. The same operators are applied, but are now ordered differently according to the values:

$\mathcal{W}_{action} =< 0.4, 0.2, 0.1, 0.9, 0.1, 0.0 >$

The calculated ordering on these revision operators is also shown in Tables A.1, A.2 and A.4. The **new-process** operator is now preferred to the **delete-condition** operator in dynamic selection, and **delete-condition release** is now preferred to **delete-condition fluid-flow** in static selection.

```
Current Motivations:
<knowledge,0.6,Fixed>
<action,0.4,Variable>
New value for action: .8
 New value for knowledge: .6
New motivations are:
<knowledge,0.6,Fixed>
<action,0.8,Variable>
```

Action motivations are now salient.

```
Current Motivations:
<knowledge,0.6,Fixed>
<action,0.8,Variable>

Maximum confidence: 0.9, and minimum confidence: 0.1
Importance: 0.8
Urgency: 0.8 with a limit of: 0.7

Accuracy: 0.8                    Credibility: 0.9

Enter reliability of observer: .9
Enter trustworthiness of observer: 1

Acceptance threshold: 0.8 ----> Action point: 0.16
Confidence: 0.648
```

```
Confidence exceeds action point - observations are accepted!

Enter observations (end with return alone):
> increase (amount-of solution2)
> decrease (amount-of solution1)
>
-----
Anomalous observations: theory refuted!
```

Note that the strength of the salient motivation, the action motivation is now 0.8. This raises the importance and thus the acceptance threshold accordingly, and would demand confidence of 0.8 or above. However, the *urgency* of the situation is high (higher than the urgency limit), so the action point is lowered accordingly, allowing the same evidence to be accepted here too, despite the increase in importance. The revisions are the same as before, but because of the need to construct a local temporary revision under action motivations based on evidence that has an associated confidence lower than the acceptance threshold, they are generated in a different order.

```
Searching for revisions ...

Revision #1
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
No change: fluid-flow
No change: add-solute
PROCESS: process8
Variables:   ?var-15
Individuals: contained-liquid ?var-15
Pconditions: precondition-osmosis-scenario ?var-15
Qconditions:
Relations:
Influences:  I+ (amount-of ?var-15) (a (process8-rate))
Revision Log: 8 new_process

PROCESS: process9
Variables:   ?var-16
Individuals: contained-liquid ?var-16
Pconditions: precondition-osmosis-scenario ?var-16
Qconditions:
Relations:
Influences:  I- (amount-of ?var-16) (a (process9-rate))
Revision Log: 8 new_process

---------------------
Revision #2
HISTORICALLY INCONSISTENT --- clash with: absorption_fails_scenario
No change: solution
```

```
No change: evaporation
No change: condensation
PROCESS: absorption
Revision Log: 8 delete conditions

No change: release
No change: fluid-flow
No change: add-solute
PROCESS: process8
Revision Log: 8 new_process

---------------------
Revision #3
HISTORICALLY INCONSISTENT --- clash with: release-fails-scenario
No change: solution
No change: evaporation
No change: condensation
No change: absorption
PROCESS: release
Revision Log: 8 delete conditions

No change: fluid-flow
No change: add-solute
PROCESS: process8
Revision Log: 8 new_process

---------------------
Revision #4
HISTORICALLY INCONSISTENT --- clash with: absorption_fails_scenario
No change: solution
No change: evaporation
No change: condensation
PROCESS: absorption
Revision Log: 8 delete conditions

PROCESS: release
Revision Log: 8 delete conditions

No change: fluid-flow
No change: add-solute

---------------------
Revision #5
HISTORICALLY INCONSISTENT --- clash with: fluid-flow-fails1-scenario
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
PROCESS: fluid-flow
Revision Log: 8 delete conditions

No change: add-solute

---------------------
```

```
Revision #6
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
No change: fluid-flow
No change: add-solute
PROCESS: process8
Revision Log: 8 new_process

--------------------
Revision #7
HISTORICALLY INCONSISTENT --- clash with: fluid-flow-fails1-scenario
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
PROCESS: fluid-flow
Revision Log: 8 delete conditions

No change: add-solute
```

Now, the first revision generated is one which will suffice for the current situation with no further reasoning. It comprises two new processes which characterize the phenomenon of osmosis. This would be better characterized by a single process, but that would require more reasoning.

### A.4.2 Grouping Observations

Below, we show how the *grouping* heuristic organises the search better. In many cases, grouping observations together provides a good basis for revision. This affects the ordering by attempting to generate revisions that account for groups of observations in progressively smaller groups. Again, the revisions are the same as before, but in a different order. Only a limited trace is shown. First, reasoning for knowledge, we get the following.

```
Current Motivations:
<knowledge,0.6,Fixed>
<action,0.4,Variable>

Searching for revisions ...

Revision #1
HISTORICALLY INCONSISTENT --- clash with: fluid-flow-fails1-scenario
No change: solution
No change: evaporation
```

```
No change: condensation
No change: absorption
No change: release
PROCESS: fluid-flow
Variables:   ?source
             ?destination
             ?path
Individuals: contained-fluid ?source
             contained-fluid ?destination
             path ?path
Pconditions: path-connection ?source ?destination ?path
Qconditions:
Relations:   Q+ (fluid-flow-rate) (pressure ?source)
             Q- (fluid-flow-rate) (pressure ?destination)
Influences:  I+ (amount-of ?destination) (a (fluid-flow-rate))
             I- (amount-of ?source) (a (fluid-flow-rate))
Revision Log: 8 delete conditions

No change: add-solute

---------------------
Revision #2
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
No change: fluid-flow
No change: add-solute
PROCESS: process8
Variables:   ?var-15
             ?var-16
Individuals: contained-liquid ?var-15
             contained-liquid ?var-16
Pconditions: precondition-osmosis-scenario ?var-15 ?var-16
Qconditions:
Relations:
Influences:  I+ (amount-of ?var-15) (a (process8-rate))
             I- (amount-of ?var-16) (a (process8-rate))
Revision Log: 8 new_process

---------------------
Revision #3
HISTORICALLY INCONSISTENT --- clash with: fluid-flow-fails1-scenario
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
PROCESS: fluid-flow
Revision Log: 8 delete conditions

No change: add-solute

---------------------
```

```
Revision #4
HISTORICALLY INCONSISTENT --- clash with: absorption_fails_scenario
No change: solution
No change: evaporation
No change: condensation
PROCESS: absorption
Revision Log: 8 delete conditions

PROCESS: release
Revision Log: 8 delete conditions

No change: fluid-flow
No change: add-solute

---------------------
Revision #5
HISTORICALLY INCONSISTENT --- clash with: release-fails-scenario
No change: solution
No change: evaporation
No change: condensation
No change: absorption
PROCESS: release
Revision Log: 8 delete conditions

No change: fluid-flow
No change: add-solute
PROCESS: process8
Revision Log: 8 new_process

---------------------
Revision #6
HISTORICALLY INCONSISTENT --- clash with: absorption_fails_scenario
No change: solution
No change: evaporation
No change: condensation
PROCESS: absorption
Revision Log: 8 delete conditions

No change: release
No change: fluid-flow
No change: add-solute
PROCESS: process8
Revision Log: 8 new_process

---------------------
Revision #7
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
No change: fluid-flow
No change: add-solute
PROCESS: process8
Revision Log: 8 new_process
```

```
PROCESS: process9
Revision Log: 8 new_process
```

The *grouping* heuristic thus allows us to consider more conservative revisions earlier, so that the consistent revision is now the second one to be considered. This contrasts with reasoning for action below in which the consistent revision is the first. This time, however, the first acceptable revision is one which represents osmosis as a single process.

```
Current Motivations:
<knowledge,0.6,Fixed>
<action,0.8,Variable>

Searching for revisions ...

Revision #1
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
No change: fluid-flow
No change: add-solute
PROCESS: process8
Variables:   ?var-15
             ?var-16
Individuals: contained-liquid ?var-15
             contained-liquid ?var-16
Pconditions: precondition-osmosis-scenario ?var-15 ?var-16
Qconditions:
Relations:
Influences:  I+ (amount-of ?var-15) (a (process8-rate))
             I- (amount-of ?var-16) (a (process8-rate))
Revision Log: 8 new_process

---------------------
Revision #2
HISTORICALLY INCONSISTENT --- clash with: fluid-flow-fails1-scenario
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
PROCESS: fluid-flow
Revision Log: 8 delete conditions

No change: add-solute

---------------------
Revision #3
No change: solution
No change: evaporation
No change: condensation
```

```
No change: absorption
No change: release
No change: fluid-flow
No change: add-solute
PROCESS: process8
Revision Log: 8 new_process

PROCESS: process9
Revision Log: 8 new_process

---------------------
Revision #4
HISTORICALLY INCONSISTENT --- clash with: absorption_fails_scenario
No change: solution
No change: evaporation
No change: condensation
PROCESS: absorption
Revision Log: 8 delete conditions

No change: release
No change: fluid-flow
No change: add-solute
PROCESS: process8
Revision Log: 8 new_process

---------------------
Revision #5
HISTORICALLY INCONSISTENT --- clash with: release-fails-scenario
No change: solution
No change: evaporation
No change: condensation
No change: absorption
PROCESS: release
Revision Log: 8 delete conditions

No change: fluid-flow
No change: add-solute
PROCESS: process8
Revision Log: 8 new_process

---------------------
Revision #6
HISTORICALLY INCONSISTENT --- clash with: absorption_fails_scenario
No change: solution
No change: evaporation
No change: condensation
PROCESS: absorption
Revision Log: 8 delete conditions

PROCESS: release
Revision Log: 8 delete conditions

No change: fluid-flow
No change: add-solute
```

```
--------------------
Revision #7
HISTORICALLY INCONSISTENT --- clash with: fluid-flow-fails1-scenario
No change: solution
No change: evaporation
No change: condensation
No change: absorption
No change: release
PROCESS: fluid-flow
Revision Log: 8 delete conditions

No change: add-solute
```

We can see that action motivations allow a quick revision to be generated that suffices for the current needs. Without grouping observations, the first consistent revision involves two applications of the **new-process** operator, resulting in two new processes which each account for one of the observations. Together, they provide an acceptable local revision. When grouping observations, only a single new process is generated, and this is in fact the desired revision corresponding to the process of *osmosis*. Under knowledge motivations, a longer path is taken that involves generating and checking five revisions without grouping, and two revisions with grouping. Thus we see that the *grouping* heuristic can significantly shorten the search if groups of observations are caused by a single process. Secondly, we see that reasoning under action motivations, though a more risky strategy because it ignores issues of consistency, and allows reasoning based on poor evidence, can generate solution states in a much smaller amount of time. The results of the four revisions are summarized in Table A.5.

| | Knowledge | Action | First consistent revision |
|---|---|---|---|
| With Grouping | 2 | 1 | one **new-process** |
| Without Grouping | 5 | 1 | two **new-processes** |

Table A.5: Number of revisions explored until consistency.

# Appendix B

# The MID Program

MID is implemented under Unix on a Sun4 in the functional language, Miranda[1]. It comprises over 1200 lines of code.

## QPT representation

Below we give details of the representation (as a variant of QPT) used in the implementation of MID.

```
Some functions to manipulate quantities where a quantity is a string
comprising an amount and an object.
If quantity = "amount-of liquid2", then qty_amt(quantity) = "amount-of"
                                    and qty_obj(quantity) = "liquid2"

> qty_amt :: quantity -> amount
> qty_amt qty = get_word (strip qty)

> qty_obj :: quantity -> object
> qty_obj qty = tl(after_word (strip qty))

RULEs are used to encode the backround knowledge rule base (BKRB)
comprising classificatory information. A rule is made up of a left
hand side (antecedent) and a right hand side (consequent), both of
which are predicates. The left hand side is said to imply the right
hand side. Abstract data type for rules:

> abstype rule
> with l_side :: rule -> predicate
>      r_side :: rule -> predicate
>      make_rule :: predicate -> predicate -> rule
>      get_rule :: (predicate, predicate) -> rule
> rule == (predicate,predicate)                || define a rule
> l_side (as,b) = as                           || get the antecedent part
```

[1]Miranda is a trademark of Research Software Ltd.

```
> r_side (as,b) = b                                  || get the consequent part
> make_rule as b = (as,b)                            || convert two predicates
>                                                    ||  (ante and cons) into a rule
> get_rule (as,b) = (as,b)                           || convert a tuple of
>                                                    ||  two predicates into a rule
```

ADT for OBSERVATIONs of the form ("increase", "amount-of liquid2").

```
> abstype observation
> with o_dir :: observation -> direction
>      o_qty :: observation -> quantity
>      get_obs :: [char] -> observation
>      make_obs :: (direction, quantity) -> observation
>      unmake_obs :: observation -> (direction, quantity)
>      show_obs :: observation -> [char]
>      test_obs :: observation
> observation == (direction, quantity)               || define an observation
> o_dir (a,b) = a                                    || get the direction part
> o_qty (a,b) = b                                    || get the quantity part
> get_obs chs = ((get_word chs), (tl (after_word chs)))
>                                                    || convert a string to an obs
> make_obs (a,b) = (a,b)                             || convert a tuple to an obs
> unmake_obs (a,b) = (a,b)                           || convert an obs to a tuple
> show_obs (a,b) = a ++ sp ++ b                      || show an observation
```

Abstract data types for elements of a process.

An INDIVIDUAL is a variable that is specified by a type predicate.

```
> abstype individual
> with i_pred :: individual -> predicate
>      i_var :: individual -> variable
>      make_ind :: (predicate, variable) -> individual
>      unmake_ind :: individual -> (predicate, variable)
>      bind_ind :: individual -> variable -> individual
>      show_ind :: individual -> [char]
> individual == (predicate, variable)                || define an individual
> i_pred (a,b) = a                                   || get the predicate part
> i_var (a,b) = b                                    || get the variable part
> make_ind (a,b) = (a,b)                             || convert a tuple into an ind
> unmake_ind (a,b) = (a,b)                           || convert an ind into a tuple
> bind_ind (a,b) c = (a,c)                           || bind a var to a predicate
>                                                    ||   to create an individual
> show_ind (a,b) = a ++ sp ++ b                      || display an individual
```

A PCONDITION is a precondition made up of a predicate and a list of variables.

```
> abstype pcondition
> with p_pred :: pcondition -> predicate
>      p_vars :: pcondition -> [variable]
>      make_p :: (predicate,[variable]) -> pcondition
>      unmake_p :: pcondition -> (predicate,[variable])
>      bind_p :: pcondition -> [variable] -> pcondition
>      show_p :: pcondition -> [char]
> pcondition == (predicate, [variable])              || define a precondition
```

```
> p_pred (a,bs) = a                                  || get the predicate part
> p_vars (a,bs) = bs                                 || get the (variable list) part
> make_p (a,bs) = (a,bs)                             || convert a tuple into a pcond
> unmake_p (a,bs) = (a,bs)                           || convert a pcond into a tuple
> bind_p (a,bs) cs = (a,cs)                          || bind a var list to a
>                                                    || predicate to create a pcond
> show_p (a,bs) = a ++ sp ++ (sp_concat bs)          || display a precondition
```

A QCONDITION is a quantity condition. It is a condition that specifies relationships between quantities. It is made up of a predicate and two quantities.

```
> abstype qcondition
> with q_pred :: qcondition -> predicate
>      q_qty1 :: qcondition -> quantity
>      q_qty2 :: qcondition -> quantity
>      make_q :: (predicate, quantity, quantity) -> qcondition
>      unmake_q :: qcondition -> (predicate, quantity, quantity)
>      show_q :: qcondition -> [char]
> qcondition == (predicate, quantity, quantity) || define a quantity condition
> q_pred (a,b,c) = a                                 || get the predicate part
> q_qty1 (a,b,c) = b                                 || get the first quantity
> q_qty2 (a,b,c) = c                                 || get the second quantity
> make_q (a,b,c) = (a,b,c)                           || convert a tuple into a qcond
> unmake_q (a,b,c) = (a,b,c)                         || convert a qcond into a tuple
> show_q (a,b,c) = a ++ sp ++ b ++ sp ++ c           || display a qcondition
```

A RELATION specifies an indirect effect. It is made up of an effect direction and two quantities, the first of which depends on the second.

```
> abstype relation
> with r_dir :: relation -> eff_dir
>      r_qty1 :: relation -> quantity
>      r_qty2 :: relation -> quantity
>      make_r :: (eff_dir, quantity, quantity) -> relation
>      unmake_r :: relation -> (eff_dir, quantity, quantity)
>      show_r :: relation -> [char]
> relation == (eff_dir, quantity, quantity)          || define a relation
> r_dir (a,b,c) = a                                  || get the direction part
> r_qty1 (a,b,c) = b                                 || get the first quantity
> r_qty2 (a,b,c) = c                                 || get the second quantity
> make_r (a,b,c) = (a,b,c)                           || convert a tuple into a rel
> unmake_r (a,b,c) = (a,b,c)                         || convert a rel into a tuple
> show_r (a,b,c) = a ++ sp ++ b ++ sp ++ c           || display a relation
```

An INFLUENCE is a direct effect of a process.

```
> abstype influence
> with i_dir :: influence -> eff_dir
>      i_qty1 :: influence -> quantity
>      i_qty2 :: influence -> quantity
>      make_inf :: (eff_dir, quantity, quantity) -> influence
>      unmake_inf :: influence -> (eff_dir, quantity, quantity )
>      show_inf :: influence -> [char]
```

```
> influence == (eff_dir, quantity, quantity)    || define an influence
> i_dir (a,b,c) = a                             || get the direction part
> i_qty1 (a,b,c) = b                            || get the first quantity
> i_qty2 (a,b,c) = c                            || get the second quantity
> make_inf (a,b,c) = (a,b,c)                    || convert a tuple into an inf
> unmake_inf (a,b,c) = (a,b,c)                  || convert an inf into a tuple
> show_inf (a,b,c) = a ++ sp ++ b ++ sp ++ c    || display an influence
```

Now we can define a PROCESS.

```
> abstype process
> with p_name :: process -> name                || get elements of process
>      vars :: process -> [variable]
>      p_inds :: process -> [individual]
>      rt :: process -> rate
>      pconds :: process -> [pcondition]
>      qconds :: process ->  [qcondition]
>      rels :: process -> [relation]
>      infs :: process -> [influence]
>      revlog :: process -> [[char]]

>      put_name :: process -> name -> process   || modify process
>      put_inf :: process -> [influence] -> process
>      put_rel :: process -> [relation] -> process
>      put_ps :: process -> [pcondition] -> process
>      put_qs :: process -> [qcondition] -> process
>      put_ind :: process -> [individual] -> process
>      put_var :: process -> [variable] -> process

>      add_to_log :: process -> [char] -> process

>      get_proc :: ([char], [variable], [(predicate, variable)], rate,
>                   [(predicate, [variable])], [(predicate, quantity,
>                   quantity)], [(eff_dir, quantity, quantity)], [(eff_dir,
>                   quantity, quantity)], [[char]]) -> process
>      put_proc :: process -> ([char], [variable], [(predicate, variable)],
>                   rate, [(predicate, [variable])], [(predicate, quantity,
>                   quantity)], [(eff_dir, quantity, quantity)],
>                   [(eff_dir, quantity, quantity)], [[char]])
>      mod_proc :: (name, [variable], [individual], rate, [pcondition],
>                   [qcondition], [relation], [influence], [[char]]) -> process

>      null_proc :: process                     || empty processes
>      show_proc :: process -> [char]           || show a process

> process == (name, [variable], [individual], rate, [pcondition],
>                   [qcondition], [relation], [influence], [[char]])
```

The following functions get each of the elements of the process.

```
> p_name (a,b,c,d,e,f,g,h,i) = a
> vars (a,b,c,d,e,f,g,h,i) = b
> p_inds (a,b,c,d,e,f,g,h,i) = c
> rt (a,b,c,d,e,f,g,h,i) = d
> pconds (a,b,c,d,e,f,g,h,i) = e
```

```
> qconds (a,b,c,d,e,f,g,h,i) = f
> rels (a,b,c,d,e,f,g,h,i) = g
> infs (a,b,c,d,e,f,g,h,i) = h
> revlog (a,b,c,d,e,f,g,h,i) = i

The following functions modify each element of the process.

> put_name (a,b,c,d,e,f,g,h,i) x = (x,b,c,d,e,f,g,h,i)
> put_inf (a,b,c,d,e,f,g,h,i) xs = (a,b,c,d,e,f,g,xs,i)
> put_rel (a,b,c,d,e,f,g,h,i) xs = (a,b,c,d,e,f,xs,h,i)
> put_ps (a,b,c,d,e,f,g,h,i) xs = (a,b,c,d,xs,f,g,h,i)
> put_qs (a,b,c,d,e,f,g,h,i) xs = (a,b,c,d,e,xs,g,h,i)
> put_ind (a,b,c,d,e,f,g,h,i) xs = (a,b,xs,d,e,f,g,h,i)
> put_var (a,b,c,d,e,f,g,h,i) xs = (a,xs,c,d,e,f,g,h,i)

> add_to_log (a,b,c,d,e,f,g,h,i) x                    || add an entry to the
>              = (a,b,c,d,e,f,g,h,i++[x])             || revision log

> get_proc (a,b,c,d,e,f,g,h,i)
>          = (a, b, (map make_ind c), d, (map make_p e), (map make_q f),
>                   (map make_r g), (map make_inf h), i)|| strings to proc
> put_proc (a,b,c,d,e,f,g,h,i)
>          = (a, b, (map unmake_ind c), d, (map unmake_p e), (map unmake_q f),
>             (map unmake_r g),(map unmake_inf h),i)    ||proc to string tuple
> mod_proc (a,b,c,d,e,f,g,h,i) = (a,b,c,d,e,f,g,h,i)    ||convert tuple to proc
> null_proc = ([],[],[],[],[],[],[],[],[])              ||empty process

> show_proc p = "PROCESS: " ++ (p_name p) ++ nl ++      || display a process
>               show_tag_list "Variables:   " (vars p) ++
>               show_tag_list "Individuals: " (map show_ind (p_inds p)) ++
>               show_tag_list "Pconditions: " (map show_p (pconds p)) ++
>               show_tag_list "Qconditions: " (map show_q (qconds p)) ++
>               show_tag_list "Relations:   " (map show_r (rels p)) ++
>               show_tag_list "Influences:  " (map show_inf (infs p)) ++
>               show_tag_list "Revision Log: " (map take_col (revlog p)) ++ nl
> where take_col x = (takewhile (~=':') x) ++ (takewhile
>                                        (~=':') (tl (dropwhile (~=':') x)))

Abstract data type for SCENARIOS.

> p_fact == pcondition                                  || some definitions
> q_fact == qcondition

> abstype scenario
> with s_name :: scenario -> name
>      s_vars :: scenario -> [variable]
>      p_facts :: scenario -> [p_fact]
>      q_facts :: scenario -> [q_fact]
>      add_p_fact :: scenario -> p_fact -> scenario
>      null_scen :: scenario
>      get_scen :: (name, [variable], [(predicate, [variable])],
>                         [(predicate, quantity, quantity)]) -> scenario
>      put_scen :: scenario -> (name, [variable], [(predicate, [variable])],
>                                    [(predicate, quantity, quantity)])
>      show_scen :: scenario -> [char]
```

```
> scenario == (name, [variable], [p_fact], [q_fact])    || define a scenario

> null_scen = ([],[],[],[]) || an empty scenario
> get_scen (a,b,c,d) = (a,b,(map make_p c),(map make_q d))
>                                    || convert a tuple into a scenario
> put_scen (a,b,c,d) = (a,b,(map unmake_p c),(map unmake_q d))
>                                    || convert a scenario into a tuple
> s_name (a,b,c,d) = a                                  || get scenario name
> s_vars (a,b,c,d) = b                                  || get the variables
> p_facts (a,b,c,d) = c                || get the pcondition facts
> q_facts (a,b,c,d) = d                || get the qcondition facts

> add_p_fact (a,b,c,d) p_f = (a, b, (c ++ [p_f]), d) || add a pcond fact
>                                    || to scenario (for characterization)

> show_scen (a,b,c,d) = "SCENARIO: " ++ a ++ nl ++      || display a scenario
> show_tag_list "Individuals: " (map show b) ++
> show_tag_list "Pfacts: " (map show_p c) ++
> show_tag_list "Qfacts: " (map show_q d)
```

Some generally useful functions that manipulate processes and scenarios:

Display an entire domain theory:

```
> show_domain :: domain -> [char]
> show_domain [] = nl
> show_domain (p:ps) = nl ++ show_proc p ++ show_domain ps
```

Display a modified domain theory.
Show modified processes completely, but only name unchanged ones:

```
> show_altered_domain :: domain -> domain -> [char]
> show_altered_domain ps xs =  show_altered_domain ps (init xs) ++
>                              show_proc (last xs), #ps ~= #xs
>                                    || ps is original domain
> show_altered_domain ps xs =  show_altered_2 ps xs, otherwise

> show_altered_2 :: domain -> [process] -> [char]
> show_altered_2 ps [] = []
> show_altered_2 ps (x:xs) = show_proc x ++ show_altered_2 ps xs, oldp ~= x
>                                                       || modified process
>                          = "No change: " ++ (p_name x) ++ nl ++ || unchanged
>                            show_altered_2 ps xs, otherwise      || process
>                                where oldp = hd (filter ok ps)
>                                      ok z = p_name z = p_name x
```

# Bibliography


[1] P. Achinstein. The method of hypothesis: What is it supposed to do, and can it do it? In P. Achinstein and O. Hannaway, editors, *Observation, Experiment, and Hypothesis in Modern Physical Science*. MIT Press/Bradford Books, Cambridge, MA, 1985.

[2] R. A. Brooks. Intelligence without reason. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 569–595, Sydney, 1991.

[3] R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

[4] C. R. Callen. Explanatory coherence and fact-finding. *Behavioral and Brain Sciences*, 14(4):739–740, 1991.

[5] A. F. Chalmers. *What is this thing called Science?* Open University Press, Milton Keynes, second edition, 1982.

[6] P. C-H. Cheng. *Modelling Scientific Discovery*. PhD thesis, Human Cognition Research Laboratory, The Open University, Milton Keynes, 1990.

[7] P. C-H. Cheng. Modelling experiments in scientific discovery. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 739–744, Sydney, 1991.

[8] P. C-H. Cheng. Approaches, models and issues in computational scientific discovery. In M. Keane and K. Gilhooly, editors, *Advances in the Psychology of thinking*, volume 1. Harvester Wheatsheaf, Hemel Hempstead, 1992.

[9] P. C-H. Cheng. Diagrammatic reasoning in scientific discovery: Modelling Galileo's kinematic diagrams. In *Working Notes of the AAAI Spring Symposium Series: Reasoning with diagrammatic representations*, Stanford, 1992.

[10] P. C-H. Cheng and H. A. Simon. The right representation for discovery: Finding the conservation of momentum. In *Proceedings of the Ninth International Conference on Machine Learning*, pages 62–71, Aberdeen, 1992.

[11] R. Cordeschi. A few words on representation and meaning. *International Studies in the Philosophy of Science*, 6(1), 1992.

[12] R. Creath. The pragmatics of observation. In A. Fine and J. Leplin, editors, *PSA 1988*, volume 1, pages 149–153, East Lansing, MI, 1988. Philosophy of Science Association.

[13] F. Crick. *What Mad Pursuit: A Personal View of Scientific Discovery.* Basic Books, New York, NY, 1988.

[14] S. Dasgupta. *Design Theory and Computer Science: Processes and Methodology of Computer Systems Design.* Cambridge University Press, Cambridge, 1991.

[15] M. desJardins. Goal-driven learning: A decision-theoretic model for deciding what to learn next. In *Proceedings of the ML92 Workshop on Machine Discovery*, pages 147–151, Aberdeen, 1992.

[16] M. E. desJardins. *PAGODA: A Model for Autonomous Learning in Probabilistic Domains.* PhD thesis, Computer Science Division, University of California, Berkeley, CA, 1992.

[17] M. d'Inverno. Using motivation as the control strategy for an autonomous agent. Unpublished Manuscript, Department of Computer Science, University College London, 1990.

[18] T. Ellman. Explanation-based learning: A survey of programs and perspectives. *ACM Computing Surveys*, 21(2):163–221, 1989.

[19] B. Falkenhainer and S. Rajamoney. The interdependencies of theory formation, revision, and experimentation. In *Proceedings of the Fifth International Conference on Machine Learning*, pages 353–366, Ann Arbor, MI, 1988.

[20] A. Fay, D. Klahr, and K. Dunbar. Are there developmental milestones in scientific reasoning? In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 333–339, Cambridge, MA, 1990.

[21] A. Flew, editor. *A Dictionary of Philosophy.* Pan Books, London, second edition, 1983.

[22] K. D. Forbus. Qualitative process theory. *Artificial Intelligence*, 24:85–168, 1984.

[23] K. D. Forbus. Qualitative physics: Past, present, and future. In E. Shrobe, editor, *Exploring Artificial Intelligence: Survey Talks from the National Conferences on Artificial Intelligence*, pages 239–296. Morgan Kaufmann, San Mateo, CA, 1988.

[24] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. In *Knowledge Discovery in Databases*, pages 1–27. AAAI Press / MIT Presss, 1991.

[25] B. R. Gaines. An overview of knowledge-acquisition and transfer. *International Journal of Man-machine Studies*, 26:453–472, 1987.

[26] R. Garigliano, A. Bokma, and D. Long. A model for learning by source control. In B. Bouchon, L. Saitta, and R. R. Yager, editors, *Lecture Notes in Computer Science, 313.* Springer-Verlag, 1988.

[27] D. A. Gillies. Comments on 'Scientific discovery as problem solving' by Herbert A. Simon. *International Studies in the Philosophy of Science*, 6(1), 1992.

[28] A. Ginsberg. *Automatic Refinement of Expert System Knowledge Bases.* Pitman Press and Morgan Kaufmann, London and San Mateo, CA, 1988.

[29] A. Ginsberg. Theory revision via prior operationalization. In *Proceedings of The Seventh National Conference on Artificial Intelligence*, pages 590–595, Saint Paul, MN, 1988.

[30] A. Ginsberg. Knowledge base refinement and theory revision. In *Proceedings of the Sixth International Conference on Machine Learning*, pages 260–265, Ithaca, NY, 1989.

[31] A. Ginsberg, W. Weiss, and P. Politakis. Automatic knowledge base refinement for classification systems. *Artificial Intelligence*, 35:197–226, 1988.

[32] N. Goodman. *Fact, Fiction and Forecast.* Bobbs-Merrill, Indianapolis, IN., 1965.

[33] M. E. Gorman. Experimental simulations of falsification. In M. Keane and K. Gilhooly, editors, *Advances in the Psychology of thinking*, volume 1. Harvester Wheatsheaf, Hemel Hempstead, 1992.

[34] I. Hacking. *Representing and Intervening.* Cambridge University Press, Cambridge, 1983.

[35] J. R. P. Halperin. Machine motivation. In J. A. Meyer and S.W. Wilson, editors, *Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats*, pages 238–246. MIT Press/Bradford Books, 1991.

[36] G. H. Harman. The inference to the best explanation. *Philosophical Review*, 74:88–95, 1965.

[37] C. G. Hempel. *Aspects of Scientific Explanation.* The Free Press, New York, NY, 1965.

[38] C. G. Hempel. *Philosophy of Natural Science.* Prentice-Hall, Englewood Cliffs, NJ, 1966.

[39] M. Hesse. Comment on Herbert Simon, "Scientific discovery as problem solving". *International Studies in the Philosophy of Science*, 6(1), 1992.

[40] R. A. Hinde. *Ethology: Its nature and relations with other sciences.* Fontana Press, 1982.

[41] J. H. Holland, K. J. Holyoak, R. E. Nisbett, and P. R. Thagard. *Induction: Processes of Inference, Learning, and Discovery.* MIT Press, Cambridge, Mass., 1986.

[42] G. Hon. Towards a typology of errors: An epistemological view. *Studies in History and Philosophy of Science*, 20(4):469–504, 1989.

[43] G. Hon. Can the monster *errour* be slain? error — a deep rooted feature of the method of experimentation that has been ignored. *International Studies in the Philosophy of Science*, 5(3):257–268, 1991.

[44] D. Hume. *The Philosophical Works of David Hume.* Black and Tait, Edinburgh, 1826.

[45] W. Iba, J. Wogulis, and P. Langley. Trading off simplicity and coverage in incremental concept learning. In *Proceedings of the Fifth International Conference on Machine Learning*, pages 73–79, Ann Arbor, MI, 1988.

[46] P. D. Karp. Hypothesis formation as design. In J. Shrager and P. Langley, editors, *Computational Models of Scientific Discovery and Theory Formation*, pages 275–317. Morgan Kaufmann, San Mateo, CA, 1990.

[47] M. Keane. *Analogical Problem Solving.* Ellis Horwood, Chichester, 1988.

[48] D. Klahr and K. Dunbahr. Dual search space during scientific reasoning. *Cognitive Science*, 12:1–48, 1988.

[49] D. Klahr, K. Dunbar, and A. L. Fay. Designing good hypotheses to test bad hypotheses. In J. Shrager and P. Langley, editors, *Computational Models of Scientific Discovery and Theory Formation*, pages 356–402. Morgan Kaufmann, San Mateo, CA, 1990.

[50] K. Kotovsky and H. A. Simon. Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology*, 4, 1973.

[51] T. S. Kuhn. *The Structure of Scientific Revolutions.* University of Chicago Press, Chicago, second edition, 1970.

[52] T. S. Kuhn. Objectivity, value judgement, and theory choice. In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, pages 277–291. University of Chicago Press, Chicago, 1977.

[53] D. Kulkarni and H. A. Simon. The processes of scientific discovery: The strategy of experimentation. *Cognitive Science*, 12:139–175, 1988.

[54] D. Kulkarni and H. A. Simon. The role of experimentation in scientific theory revision. In *Proceedings of the Sixth International Conference on Machine Learning*, pages 278–283, Ithaca, NY, 1989.

[55] D. Kulkarni and H. A. Simon. Experimentation in machine discovery. In J. Shrager and P. Langley, editors, *Computational Models of Scientific Discovery and Theory Formation*, pages 255–273. Morgan Kaufmann, San Mateo, CA, 1990.

[56] Z. Kunda. The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498, 1990.

[57] I. Lakatos. *Proofs and Refutations: The Logic of Mathematical Discovery.* Cambridge University Press, Cambridge, 1976.

[58] D. Lamb. *Discovery, Creativity and Problem-Solving.* Avebury, Aldershot, 1991.

[59] P. Langley, G. L. Bradshaw, and H. A. Simon. Rediscovering chemistry with the BACON system. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach*, pages 307–329. Tioga, Palo Alto, 1983.

[60] P. Langley, H. A. Simon, G. L. Bradshaw, and J. M. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Processes.* MIT Press, Cambridge, Mass., 1987.

[61] P. Langley, J. M. Zytkow, H. A. Simon, and G. L. Bradshaw. The search for regularity: four aspects of scientific discovery. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach, Volume II*, pages 425–469. Morgan Kaufmann, Los Altos, CA, 1986.

[62] D. Leake. *Evaluating Explanations.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1992.

[63] D. Leake and A. Ram. Goal-drivern learning: Fundamental issues and symposium report. Technical Report 85, Cognitive Science Program, Indiana University, Bloomington, Indiana, 1993.

[64] D. B. Leake. Evaluating explanations. In *Proceedings of The Seventh National Conference on Artificial Intelligence*, pages 251–255, 1988.

[65] D. B. Leake. Task-based criteria for judging explanations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 325–332, Cambridge, MA, 1990.

[66] D. B. Leake. Goal-based explanation evaluation. *Cognitive Science*, 15:509–545, 1991.

[67] R. Lelouche and S. Doublait. Qualitative reasoning with bluff and beliefs in a multi-actor environment. *International Journal of Man-Machine Studies*, 36:149–165, 1992.

[68] D. B. Lenat and E. A. Feigenbaum. On the thresholds of knowledge. *Artificial Intelligence*, 47:185–250, 1991.

[69] D.B. Lenat, R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd. Cyc: Towards programs with common sense. *Communications of the ACM*, 33(8), 1990.

[70] I. Levi. *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability and Chance.* MIT Press, Cambridge, MA, 1980.

[71] I. Levi. Truth, fallibility and the growth of knowledge. In *Decisions and revisions: Philosophical essays on knowledge and value*, pages 109–127. Cambridge Unversity Press, Cambridge, 1984.

[72] M. Luck. A six-stage model for inductive reasoning. Research Note RN/90/14, Department of Computer Science, University College London, 1990.

[73] M. Luck. Motivations in inductive discovery: Reasoning for knowledge and action. Research Note RN/92/80, Department of Computer Science, University College London, 1992.

[74] M. Luck. Nuggets of information. *Computing*, pages 34–35, 22 October, 1992.

[75] P. Maes. The dynamics of action selection. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 991–997, Detroit, 1989.

[76] P. Maes. How to do the right thing. *Connection Science*, 1(3):291–323, 1989.

[77] P. Maes. A bottom-up mechanism for behaviour selection in an artificial creature. In J. A. Meyer and S.W. Wilson, editors, *Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats*, pages 238–246. MIT Press/Bradford Books, 1991.

[78] S. Marsh. Trust and reliance in multi-agent systems: A preliminary report. In *Pre-Proceedings of the Fourth European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, 1992.

[79] J. W. McAllister. Truth and beauty in scientific reason. *Synthese*, 78:25–51, 1989.

[80] T. M. Mitchell, R. Keller, and S. Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine Learning*, 1:47–80, 1986.

[81] R. Oehlmann, D. Sleeman, and P. Edwards. Self-questioning and experimentation in an exploratory discovery system. In *Proceedings of the ML92 Workshop on Machine Discovery*, pages 41–50, Aberdeen, 1992.

[82] D. Oldroyd. *The Arch of Knowledge.* Methuen and Co., New York, 1986.

[83] P. O'Rorke. Coherence and abduction. *Behavioural and Brain Sciences*, 12(3):484, 1989.

[84] P. O'Rorke, S. Morris, and D. Schulenburg. Theory formation by abduction: A case study based on the chemical revolution. In J. Shrager and P. Langley, editors, *Computational Models of Scientific Discovery and Theory Formation*, pages 197–224. Morgan Kaufmann, San Mateo, CA, 1990.

[85] J. Pearl. *Heuristics: Intelligent search strategies for computer problem solving.* Addison-Wesley, Reading, MA, 1984.

[86] D. Poole. Hypo-deductive reasoning for abduction, default reasoning and design. In *Working Notes of the AAAI Spring Symposium on Automated Abduction (Technical Report 90-32)*, pages 106–110, Irvine: University of California, Department of Information and Computer Science, 1990.

[87] K. R. Popper. *The Logic of Scientific Discovery.* Hutchinson, London, 1959.

[88] W. V. Quine and J. S. Ullian. *The Web of Belief.* Random House, New York, 1978.

[89] S. Rajamoney and G. Dejong. The classification, detection and handling of imperfect theory problems. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 205–207, Milan, 1987.

[90] S. A. Rajamoney. Exemplar-based theory rejection: An approach to the experience consistency problem. In *Proceedings of the Sixth International Conference on Machine Learning*, pages 284–289, Ithaca, NY, 1989.

[91] S. A. Rajamoney. *Explanation-based Theory Revision: An Approach to the Problems of Incomplete and Incorrect Theories.* PhD thesis, Department of Computer Science, University of Illinois, Urbana, Illinois, 1989.

[92] S. A. Rajamoney. A computational approach to theory revision. In J. Shrager and P. Langley, editors, *Computational Models of Scientific Discovery and Theory Formation*, pages 225–253. Morgan Kaufmann, San Mateo, CA, 1990.

[93] S. A. Rajamoney and G. F. Dejong. Active ambiguity reduction: An experiment design approach to tractable qualitative reasoning. Technical Report UILU-ENG-87-2225, Artificial Intelligence Research Group, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, Illinois, 1987.

[94] S. A. Rajamoney and G. F. Dejong. Active explanation reduction: An approach to the multiple explanations problem. In *Proceedings of the Fifth International Conference on Machine Learning*, Ann Arbor, 1988.

[95] A. Ram. Decision models: A theory of volitional explanation. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 198–205, Cambridge, MA, 1990.

[96] A. Ram. Goal-based explanation. In *Working Notes of the AAAI Spring Symposium on Automated Abduction (Technical Report 90-32)*, pages 26–29, Irvine: University of California, Department of Information and Computer Science, 1990.

[97] A. Ram. Knowledge goals: A theory of interestingness. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 206–214, Cambridge, MA, 1990.

[98] A. Ram and D. Leake. Evaluation of explanatory hypotheses. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pages 867–871, Chicago, IL., 1991.

[99] P. Reimann. *Problem Solving Models of Scientific Discovery Learning Processes.* Peter Lang, Frankfurt am Maim, 1990.

[100] D. Rose and P. Langley. STAHLp: Belief revision in scientific discovery. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 528–532, Philadelphia, PA., 1986.

[101] D. Rose and P. Langley. A hill-climbing approach to machine discovery. In *Proceedings of the Fifth International Conference on Machine Learning*, pages 367–373, Ann Arbor, MI, 1988.

[102] S. Russell. Inductive learning by machines. *Philosophical Studies*, 64:37–64, 1991.

[103] C. Schaffer and A. Prieditis. Evaluating machine discovery systems. In *Proceedings of the ML92 Workshop on Machine Discovery*, pages 166–167, Aberdeen, 1992.

[104] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In J. A. Meyer and S.W. Wilson, editors, *Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats*, pages 465–474. MIT Press/Bradford Books, 1991.

[105] U. Schnepf. Robot ethology: A proposal for the research into intelligent autonomous systems. In J. A. Meyer and S.W. Wilson, editors, *Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats*, pages 465–474. MIT Press/Bradford Books, 1991.

[106] Y. Shoham. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence.* MIT Press, Cambridge, MA, 1988.

[107] H. A. Simon. Motivational and emotional controls of cognition. In *Models of Thought*, pages 29–38. Yale University Press, 1979.

[108] H. A. Simon. Scientific discovery as problem solving: reply to critics. *International Studies in the Philosophy of Science*, 6(1):69–88, 1992.

[109] H. A. Simon and K. Kotovsky. Human acquisition of concepts for sequential patterns. *Psychological Review*, 70(6):534–546, 1963.

[110] H. A. Simon and G. Lea. Problem solving and rule induction: a unified view. In L. W. Gregg, editor, *Knowledge and Cognition*, pages 105–127. Lawrence Erlbaum Associates, Potomac, MD., 1974.

[111] H. A. Simon and R. E. Valdes-Perez. The evaluation and reporting of discovery systems. In *Proceedings of the ML92 Workshop on Machine Discovery*, pages 163–165, Aberdeen, 1992.

[112] D. H. Sleeman, M. K. Stacey, P. Edwards, and N. A. B. Gray. An architecture for theory-drivern discovery. In *Proceedings of the 4th European Working Session on Learning*, 1989.

[113] A. Sloman. Motives, mechanisms, and emotions. *Cognition and Emotion*, 1(3):217–233, 1987.

[114] A. Sloman and M. Croucher. Why robots will have emotions. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pages 197–202, Vancouver, B.C., 1981.

[115] P. Thagard. The best explanation: criteria for theory choice. *Journal of Philosophy*, 75:76–92, 1978.

[116] P. Thagard. Computational models in the philososphy of science. In A. Fine and P. Machamer, editors, *PSA 1986*, volume 2, pages 329–335, East Lansing, 1987. Philosophy of Science Association.

[117] P. Thagard. *Computational Philosophy of Science*. MIT Press/Bradford Books, Cambridge, Mass., 1988.

[118] P. Thagard. Explanatory coherence. *Behavioural and Brain Sciences*, 12(3):435–502, 1989.

[119] P. Thagard. Extending explanatory coherence. *Behavioural and Brain Sciences*, 12(3):490–499, 1989.

[120] P. Thagard. Explanatory coherence and naturalistic decision making. In *Working Notes of the AAAI Spring Symposium on Automated Abduction (Technical Report 90-32)*, pages 125–129, Irvine: University of California, Department of Information and Computer Science, 1990.

[121] P. Thagard. Defending explanatory coherence. *Behavioural and Brain Sciences*, 14(4):745–748, 1991.

[122] P. Thagard. Adversarial problem solving: Modelling an opponent using explanatory coherence. *Cognitive Science*, 16:123–149, 1992.

[123] P. Thagard. *Conceptual Revolutions*. Princeton University Press, Princeton, NJ, 1992.

[124] P. Thagard and Z. Kunda. Hot cognition: Mechanisms for motivated inference. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, pages 753–763, Hillsdale, NJ, 1987. Lawrence Erlbaum Associates.

[125] P. Thagard and G. Nowak. The explanatory coherence of continental drift. In A. Fine and J. Leplin, editors, *PSA 1988*, volume 1, pages 118–126, East Lansing, MI, 1988. Philosophy of Science Association.

[126] P. Thagard and G. Nowak. The conceptual structure of the geological revolution. In J. Shrager and P. Langley, editors, *Computational Models of Scientific Discovery and Theory Formation*, pages 27–72. Morgan Kaufmann, San Mateo, CA, 1990.

[127] M. Thost. Generating facts from opinions with information source models. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 531–536, Detroit, 1989.

[128] F. Toates and P. Jensen. Ethological and psychological models of motivation — towards a synthesis. In J. A. Meyer and S.W. Wilson, editors, *Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats*, pages 194–205. MIT Press/Bradford Books, 1991.

[129] P. E. Utgoff. Shift of bias for inductive concept learning. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach, Volume II.* Morgan Kaufmann, Los Altos, CA, 1986.

[130] M. G. Walker. How feasible is automated discovery? *IEEE Expert*, 2(1):69–82, 1987.

[131] D. L. Waltz. Eight principles for building an intelligent robot. In J. A. Meyer and S.W. Wilson, editors, *Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats.* MIT Press/Bradford Books, 1991.

[132] P. C. Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12:129–140, 1960.

[133] J. Watkins. Has BACON vindicated Kant? *International Studies in the Philosophy of Science*, 6(1), 1992.

[134] J. Wolstencroft. Restructuring, reminding and repair: What's missing from models of analogy? *AI Communications*, 2:103–118, 1989.

[135] R. Zembowicz and J. M. Zytkow. Automated discovery of empirical equations form data. In *Methodologies for Intelligent Systems: Sixth International Symposium, ISMIS '91, (LNAI 542)*, pages 429–440, Charlotte, NC., 1991.

[136] R. Zembowicz and J. M. Zytkow. Discovery of regularities in databases. In *Proceedings of the ML92 Workshop on Machine Discovery*, pages 18–27, Aberdeen, 1992.

[137] J. Ziman. *Reliable Knowledge: An Exploration of the Grounds for Belief in Science.* Cambridge University Press, Cambridge, 1978.

[138] J. Zytkow and H. A. Simon. Normative systems of discovery and logic of search. *Synthese*, 74:65–90, 1988.

[139] J. Zytkow, J. Zhu, and R. Zembowicz. The first phase of real-world discovery: Determining repeatability and error of experiments. In *Proceedings of the Ninth International Conference on Machine Learning*, pages 480–485, Aberdeen, 1992.