UNIVERSITY OF ALBERTA

Statistically Motivated Defaults

by

Scott D. Goodwin

A thesis
submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree
of Doctor of Philosophy

Department of Computing Science

Edmonton, Alberta
Fall 1991

ISBN 0-315-70139-0

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Scott D. Goodwin

TITLE OF THESIS: Statistically Motivated Defaults

DEGREE: Doctor of Philosophy

YEAR THIS DEGREE GRANTED: 1991

(Signed) Scott Goodwin

Scott D. Goodwin
131 Centennial Street
Regina, Saskatchewan
Canada, S4S 6W3

Date: Oct 8/91

Sorrows come to stretch out spaces in the heart for joy.
E. Markham

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recomended to the Faculty of Graduate Studies and Research for acceptance a thesis entitled Statistically Motivated Defaults submitted by Scott D. Goodwin in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

R. Goebel

R. Goebel

J. Pelletier

R. Elio

Bernard Linsky

B. Linsky

H. Kyburg, Jr.

OCT - 3 1991

To Monarch Woods

Abstract

Default reasoning is a fundamental area of research in Artificial Intelligence. A default is knowledge that is generally true though it admits exceptions (e.g., birds fly, objects retain their colour when moved, and people with colds cough). The approach taken here is to view a default as a statistical claim about the world plus an applicability criterion. The assumed applicability of a default to a particular case is thereby governed by the same criteria governing the assumed applicability of statistical knowledge to a particular case—the problem of determining default applicability is essentially the famous problem of reference class selection in probability theory.

We examine the problem of reference class selection within the context of Bacchus and Halpern's combined probability logic and we introduce a new approach based on the idea of second order randomization. Under this scheme, the assumed independence properties of particular predicates are based on the independence properties of randomized predicates. This is a natural extension to first order randomization, where the degree of belief in a proposition about a particular individual is based on the statistical properties of randomized individuals. The flavour is much like that of maximum entropy approaches but we avoid problems such as syntax sensitivity. This approach accounts for many intuitive default inferences and solves problems involving conflicting sources of statistical knowledge that present difficulties for many other approaches.

Acknowledgements

This work would not have been possible without the significant intellectual, emotional, spiritual, and financial support I have received from my colleagues, friends, and family.

The first and largest share of thanks, I extend to my supervisor Randy Goebel. He expended much effort in building up an environment in which it is easy to be productive and creative, and in which it is safe to discuss ideas freely. Generous praise, thoughtful criticism, and careful apportioning of credit have been his hallmark. He has also been very generous with financial support over the years. He has provided funding for trips to Germany, Japan, and many other places. He has also made equipment available and provided many other facilities that were of significant benefit. Over the years, he has served as my mentor, my "big brother," and my friend. His constant striving for clarity, simplicity, and unity has inspired both my professional and personal life. As this time of my life comes to a close, I look back with a tear in my eye and a smile in my heart.

And to Bon Bon and Chum Chum, for Monarch Woods and Marshmallow Cloud Dreams—the line is thinly drawn 'tween joy and sorrow.

To Carolyn Morris, for being a better friend than I.

To André Trudel for Hearts and Flowers, and Face Plant. To Sam and Max, for the cuddles. To Annie, in appreciation for your special attention. (To Cindy, so I don't get heck.)

To Eric Neufeld, for the drinkii g money.

To Bruce Spencer, for listening when I really needed it.

To Qiang Yang and the baby ducks.

To Dave Gillies, for the Human Wave Attack and Infinite Machine Guns and Rockets Lost in Space.

To Keith Mah, for Tuesday night at the movies.

Special thanks to Fahiem Bacchus, for the job (when I ran away to home) and for support for research.

To Robin Cohen, for carefully reading my dissertation when no one else could make time.

To Kari, Jodi, and Kelly, for eating all the lollypops.

To my parents, Frank and Colleen, and my sisters: Cathy, Vicki, Janice, Kelly, Lynn, and Tracy. For believing in me, when I did not.

And finally, to my many, many friends: George Ferguson, Pat Fitzsimmons, Dekang Lin, Abdul Sattar, Bonita Wong, Ramesh Sankaranarayana, Lingyan Shu, Kirk Morris, Chrysanne DiMarco, Steven Dai, Vickie Martin, Valerie Clark, Janelle Harms, Peter Van Beek, Fei Song, Jennifer Chu, Michael (Mudcat) Gruninger, Dale Schuurmans, Howard Hamilton, Charley Kasmierski (who denies he likes to see his name in print).

And deep, deep apologies to my dear friends whom I have unforgivably neglected to mention.

Contents

List of Figures

List of Examples

Chapter 1

Introduction

1.1 Representation and Reasoning Controversy

Great controversy and confusion has arisen lately in Artificial Intelligence (AI) regarding the principles of representing and reasoning in the presence of uncertainty [38, 51]. A resurgence of the declarativism/proceduralism debate of the seventies was sparked by McDermott's disillusionment with the logicist approach [49].

The logicist's credo—that knowledge can and should be represented declaratively, independently of how the knowledge is used—hinges on two premises. First, the conclusions that deductively follow from a set of premises follow regardless of how those conclusions will be used—truth is truth. Second, a significant part of reasoning is deductive. There is no question that the first premise is true since soundness is a property of deduction, but McDermott argues that the second premise is erroneous. Because plausible reasoning is not logical deduction, McDermott reluctantly suggests that AI must inevitably resort to procedural ad hocery. He reached the pessimistic conclusion that

> "...we must resign ourselves to writing programs, and viewing knowledge representations as entities to be manipulated by the programs." [49, p. 159]

While logical deduction is useful in determining what conclusions follow from a set of axioms, it says nothing about how to rationally determine and maintain a set of beliefs. According to Israel [26], what is necessary for the formalization of rational belief (and rational inference) is the specification of a set of rational epistemic policies of belief fixation and revision. Furthermore, these policies must necessarily be heuristic in nature and are akin to scientific procedures (i.e., the scientific method). Israel argued that common sense reasoning should be considered within the framework of scientific theory formation. If progress is to be made, Israel contends, we should not confine our attention to semantics and proof-theory; rather we should turn to epistemology and the philosophy of science.

More recently, the probabilists, for example, Cheeseman, have suggested that "the logic mafia" have overlooked, indeed ignored, probability theory (in its various guises) as the natural tool for plausible reasoning [6]. Cheeseman claims

> "...that the difficulties McDermott described are a result of insisting on using logic as the language of common sense reasoning.

> If, instead, (Bayesian) probability is used, none of the technical difficulties found in using logic arise." [6, p. 58]

Perhaps this is so, but as the flurry of responses to Cheeseman's position paper suggest, probability theory has its own closet full of skeletons [51].

Yet the probabilists cannot be dismissed lightly. Much recent work has suggested that even if numeric probability theory is not a panacea for all the ills of the logicist approach, it has many useful concepts like *degrees of belief*, *conditionalization*, *randomness*, and *independence* that have no direct analogue in logic (e.g., [1, 2, 12, 18, 37, 58, 55]). In view of the evidence, logicists can no longer ignore the importance of probabilistic concepts and the probabilists cannot overlook the fundamental computational importance of locality of logical entailment.

Cheeseman's position was the catalyst for the current controversy between the logicists and the probabilists and has many researchers struggling to determine the appropriate role of logic and probability in AI [51].

1.2 Motivation

This dissertation is the result of the author's own struggle to determine the appropriate role of logic and probability in AI. The work evolved from the author's previous work on hypothetical reasoning and the frame problem in temporal reasoning [15]. In that work, two kinds of knowledge were of prime concern: *defaults*—propositions that are generally true but admit exceptions, such as birds fly,—and *theory preference*—any criteria by which sets of assumptions are ranked. The work left several epistemological issues unsettled:

1. What kind of knowledge do defaults and preferences encode?
2. In what sense are defaults and preferences true or correct?
3. How are defaults and preferences determined or verified in the domain?

The motivation of this current work is to seek answers to such epistemological questions. More particularly, we seek a deeper understanding of the nature of default knowledge used in common sense reasoning and we desire a specification of defaults that provides a clear representation of domain knowledge in a principled way.

An offshoot of our primary motivation is that we seek to develop a tool to facilitate the specification and exploration of default knowledge, i.e., we would like a hypothetical reasoning system that allows us to experiment with potential specifications and various theory preferences. Hypothetical reasoning is the subject of the next section.

1.3 A Hypothetical Reasoning Framework

Since our knowledge of the world is imperfect, our representation of the world is necessarily only an approximation. Regardless of the limitations on our ability to represent the world, we still want to be able to reason about it. In order to draw useful conclusions about a domain for which our knowledge is inaccurate or incomplete, we must make assumptions. It seems that the only reasonable way to manage these assumptions so as to draw useful conclusions is to adopt a scientific theory formation approach. This approach involves building, testing, and revising theories to explain observations and make predictions. In view of our imperfect knowledge, it seems the best we can hope for is to build plausible theories and revise them when they are found defective.

Though the ability to make assumptions, draw conclusions, and revise beliefs is an essential part of rational (and common sense) reasoning, this ability alone is not enough—we must be able to compare alternative sets of possible assumptions and determine those which are most plausible. Representing our uncertain and incomplete knowledge, reasoning effectively (making plausible assumptions) in spite of our imperfect knowledge, and rationally maintaining and revising beliefs: these are all vital aspects of common sense reasoning.

The development of nonmonotonic reasoning systems was, in part, motivated by the inability of logical deduction to capture the forms of rational inference typically involved in common sense reasoning. Because logical deduction is monotonic and because rational inference is not, attempts were made to extend classical logic to allow nonmonotonic reasoning [46, 50, 75]. Israel has criticized these formalizations of nonmonotonic reasoning and argued that, instead of developing extensions to logic, nonmonotonic reasoning should be considered within the framework of scientific theory formation [26].

He contends that such an approach is not only the best we could hope for, but is also the only thing that makes sense.

In the spirit of Israel's proposal, Goebel, Poole and their colleagues have been investigating the theory formation (or hypothetical reasoning) approach to common sense reasoning in the Theorist project [70]. Theorist views reasoning as scientific theory formation (rather than as deduction). Science is concerned, not merely with collecting facts, but also with finding explanations, making predictions, testing and revising theories. Reasoning in the Theorist framework involves building theories that explain observations or make predictions.

By specifying a language of facts and assumptions and a procedure for drawing and retracting assumption-based conclusions, the Theorist framework provides a natural tool for investigating common sense reasoning.

The next section turns to the question of justifying assumptions in hypothetical reasoning. In particular, the statistical interpretation of defaults is introduced.

1.4 Statistical Interpretation of Defaults

As Cheeseman [6] noted, "the logic mafia" have ignored probability as a tool in knowledge representation.

> "...ever since McCarthy and Hayes [48] proclaimed probabilities
> to be 'epistemologically inadequate,' AI researchers have shunned

> probability adamantly. Their attitude has been expressed through commonly heard statements like 'The use of probability requires a massive amount of data,' 'The use of probability requires the enumeration of all possibilities,' and 'People are bad probability estimators.' 'We do not have those numbers,' it is often claimed, and even if we do, 'We find their use inconvenient.' " [58, p. 15]

But if one takes the view that "probability is not really about numbers: it is about the structure of reasoning," as Shafer did [58, p. 15], then the value of probability as a knowledge representation tool becomes more apparent. In particular, it is hard not to appeal to probability in the analysis and representation of plausible knowledge such as default assumptions in hypothetical reasoning.

One approach to answering the question, posed in section 1.2, about the kind of knowledge defaults encode is to propose a probability model that is alleged to capture the content of a set of facts and defaults. Probability theory can then be used to determine the plausibility of various conclusions. If some conclusion which is not plausible in the domain of interest is deemed plausible by the probability model, the various assertions in the probability model can be examined to find those that are incompatible with the domain or to determine if additional assertions are necessary. If the probability model needs to be changed then either

- the content of the facts and defaults didn't correspond to the domain knowledge; or,
- the original transformation from facts and defaults to probability assertions didn't capture their content.

If the former is true, this exercise has taught us something more about the domain (i.e., identified missing or incorrect facts and defaults). If the latter is true, we have learned something about the content of the facts and defaults (i.e., what they say about the domain).

There is much work in the literature that proposes probabilistic interpretations for defaults. Pearl's ϵ-semantics takes a default to be a conditional probability arbitrarily close to one [58]. For example, the probability assertion p(fly|bird) = 1-ϵ is taken to mean that birds fly by default.

A second approach, due to Neufeld, is to take the minimal meaning of defaults to be a "favouring." Here the birds fly default is taken to mean (at least) that knowing something is a bird *favours* concluding it flies [55]. The probability assertion p(fly|bird) > p(fly) expresses this.

A third approach, due to Bacchus, is to take a default to be a conditional probability greater than some threshold c [2]. In this case, the probability assertion p(fly|bird) > c represents that birds fly by default.

Here we adopt a view similar to that of Bacchus and interpret the Meta-Theorist statement (see section 6.1)

default birdsfly(X): fly(X) $\leftarrow(X)-$ bird(X).

as encoding the statistical knowledge that most X's that are birds also fly. Section 6.4 provides some indication of just how far this interpretation takes us toward capturing the intuitive content of defaults.

1.5 The Thesis

A fruitful approach to the specification of computational models of common sense reasoning is the development of a hypothetical reasoning framework based on scientific theory formation in which assumptions and their justifications are explicitly represented and reasoned about. Representing and reasoning in a hypothetical reasoning framework involves specifying at least the following: what is *known* to be true about the domain; what can be *assumed* about the domain; under which conditions are assumptions *justified*; and, given the above, what can be *inferred* about the domain.

While logic is useful for the specification and implementation of such a framework, some concepts from probability theory are essential and yet they are not part of the machinery of logic. The most important of these are: *degrees of belief*, *conditionalization*, *randomness*, and *independence*. The importance, for instance, of the concept of independence is revealed by its ubiquity in common sense reasoning.

The power of goal-directed methods stems directly from independence—irrelevant information need not be considered. In logic, every proposition is independent of every other proposition unless they are related via an implication—hence the context independence of modus ponens and the existence of goal-directed proof procedures. In common sense reasoning a dependency between propositions might affect the reasoning even if there is no known implication between the propositions. For example, learning that a bird lives in the Antarctic might cast doubt on its flying abilities even if living in the Antarctic does not imply the bird does not fly—such cases are not directly addressed by the machinery of logic.

In probability theory, a random variable is independent of another random variable only if its independence is asserted. Furthermore, independence must be asserted for every context in which it holds—hence the context dependence of the probability of a random variable. Probability theory suffers from a kind of qualification problem [45, p. 1040]; to say that a random variable depends only on some set of random variables, we must assert its independence from every other random variable.

Logic and probability theory can be viewed as two extremes on a continuum of varying degrees of (implicit or assumed) independence. The middle ground is more suitable than either extreme for computational specifications of common sense reasoning.

In a hypothetical reasoning framework, the appropriate place to articulate independence is at the meta-level, i.e., at the level of *justification* of assumptions. The meta-level is also the appropriate place to articulate other probability theory concepts.

Indeed the semantic difference between the various kinds of hypotheses (e.g., defaults, conjectures, etc.) arises at the meta-level. In the case of defaults, in particular, their semantics are intimately intertwined with the above concepts from probability theory.

In view of the evidence, logicists can no longer ignore the importance of probabilistic concepts and the probabilists cannot overlook the fundamental computational importance of the goal-directedness offered by logic.

The major thesis this dissertation defends is that reference class selection policies based on Reichenbach's principle (and, analogously, default inheritance based on subset preference (specificity)) are inadequate for common sense reasoning. Instead, reference class selection should be based on a principle of second order direct inference (and second order randomization) where independencies of particular predicates are inferred from independencies of the set of similar predicates.

1.6 The Contribution of this Work

This work provides a deeper understanding of the intuitive concept of default used in common sense reasoning. This understanding stems from the use of logic and probability as tools to unravel and specify the meaning of defaults. We have identified and articulated an appropriate role of logic and probability in AI, namely, the analysis and specification of common sense reasoning concepts.

The use of these tools has led to the specification of a statistically motivated semantics for defaults, which in turn facilitated the representation of domain knowledge in a way that avoids the problems of current representations—not by yet another "technical tweak"—but by having a well-founded representation.

A major contribution is the introduction of Second Order Direct Inference in Chapter 5. This provides a treatment of default inheritance that more closely mirrors intuition than do approaches based on Reichenbach's principle and other specificity approaches. This is the key difference with almost all the approaches to default reasoning that we are aware of, except perhaps maximum entropy approaches. Examples such as the Vaccinated Child problem (Example 6.9), the Russian Roulette problem (Example 7.1), and the Elephants and Zookeepers problem (Example 6.2) are typical cases not correctly handled by other approaches.

Another contribution is the extension of the basic Theorist hypothetical reasoning framework to include a meta-level which allows the specification of theory preference and pruning criteria, several kinds of hypotheses (e.g., default, convention, and conjecture), and additional kinds of inference (e.g., prediction). This Meta-Theorist framework provides a platform to experiment with potential specifications of defaults. The implementation of Meta-Theorist made it possible to provide empirical evidence to support our proposed statistical interpretation of defaults. Additionally, Meta-Theorist can be seen as a general tool for hypothetical reasoning—it is not limited to tasks involving defaults.

More specifically, the contributions are:

- the interpretation of Meta-Theorist defaults as statistical assertions:
- the specification of statistically motivated defaults in Meta-Theorist;
- the development of a reference class selection policy:
- the introduction of Second Order Direct Inference;
- the Viable Inheritance Correspondence conjecture;
- the extension of basic Theorist to Meta-Theorist;
- two theorems about incremental theory pruning and preference; and
- the implementation of Meta-Theorist.

1.7 Outline of the Presentation

The next chapter describes the basic Theorist hypothetical reasoning framework and the philosophy behind it.

An extension to Theorist is described in Chapter 3, along with two theorems that justify incrementally computing theory pruning and preference.

Chapter 4 discusses probabilistic knowledge and inference. In particular, it examines Bacchus's probability logic [3] which is adopted and extended for use as a formal tool in the analysis and specification of defaults.

Chapter 5 examines the problem of reference class selection. The reference class selection policy in Bacchus's formalism is shown to be inadequate. A new policy is proposed based on the average over the reference class to which we are indifferent.

The expressive power of Bacchus's formalism makes it difficult to implement. Chapter 6 develops a specification for statistically motivated defaults in Meta-Theorist. These defaults correspond to a restricted form of statistical assertion in Bacchus's probability logic, namely, statistical majority. Several examples that indicate the merit and shortcomings of this approach are discussed.

The final chapter discusses the significance of the current work, and possible directions for future work.

Chapter 2

Theorist

In view of our imperfect knowledge of the world, it seems the best we can hope for is to build plausible theories and revise them when they are found defective. The methodology of observe, hypothesize, predict, test, and evaluate, which has proved so fruitful in science, should also be useful when applied to prescientific common sense reasoning and rational belief. Based on a philosophy inspired by Popper [72] and in the spirit of Israel's proposal [26], Goebel, Poole and their colleagues have been investigating the scientific theory formation approach to common sense reasoning in the Theorist project [84, 70]. This chapter describes the Theorist hypothetical reasoning framework and the philosophy behind it. As the author was one of the contributors to the Theorist project, this chapter, besides being a review of previous work, presents the author's current view. In the following chapter, our extension, called Meta-Theorist, is described along with its implementation.

2.1 Philosophical Basis

Reasoning in the Theorist framework involves building theories to explain observations. In the philosophy of science, the deductive nomological model of explanation [24] defines an explanation as a valid deductive argument. The premises of the argument include nomological general statements (laws) and other singular statements (called initial conditions). These statements deductively entail the conclusion. In this sense, the conclusion is explained by the premises; that is, to explain an observation is to deduce it from a law. Scientific explanation of observations involves a subsumption argument under laws.

Laws are universal generalizations that express relations that hold between various properties. Nomic necessity and hypothetical force distinguish laws from ordinary generalizations. To say that a generalization has nomic necessity is to say that the relationship it expresses is somehow necessary. This element of necessity extends to the unobserved, the unrealized, and the hypothetical counterfactual. For example, from the law that objects denser than water sink, we can infer that if an ice cube were denser than water (which it isn't) then it would sink.

A question that has troubled philosophers since early times is: "Where do laws come from?" Given that universal laws govern the unobserved, it is clear that they cannot be arrived at deductively from experience; similarly, given that universal laws govern the unobservable, the hypothetical, and the counterfactual, it is clear that they cannot be arrived at inductively either. Rescher answers the question as follows:

> "The basic fact of the matter—and it is a fact whose importance cannot be overemphasized—is that the elements of nomic necessity and hypothetical force are not to be extracted from the evidence. They are not *discovered* on some basis of observation at all; they are *supplied.* The realm of hypothetical counterfact is inaccessible to observational or experimental examination. Lawfulness is not found in or extracted from the evidence, it is superadded to it. *Lawfulness is a matter of imputation.* When an empirical generalization is designated as a law, this epistemological status is *imputed* to it. Lawfulness is something which a generalization could not *in principle* earn entirely on the basis of warrant by the empirical facts. Men impute lawfulness to certain generalizations by according them a particular role in the epistemological scheme of things, being prepared to use them in special ways in inferential contexts (particularly hypothetical contexts), and the like.
>
> When one looks at the explicit formulation of the overt *content* of a law all one finds is a certain generalization. Its lawfulness is not a part of what the law asserts at all; it is nowhere to be seen in its overtly expressed content as a generalization. Lawfulness is not a matter of what a generalization *says*, but a matter of *how it is to be used.* By being prepared to put it to certain kinds of uses in modal and hypothetical contexts, it is *we*, the users, who accord to a generalization its lawful status thus endowing it with nomological necessity and hypothetical force. Lawfulness is thus not a matter of the assertive content of a generalization, but of its epistemic status, as determined by the ways in which it is deployed in its applications." [76, p. 107]

The key point to draw from this is: "Lawfulness is thus not a matter of the assertive content of a generalization, but of its epistemic status, as determined by the ways in which it is deployed in its applications." As we shall see, the power of Theorist derives from the special epistemic status it gives to possible hypotheses. The nonmonotonic nature of Theorist results from the ways in which possible hypotheses are used.

A distinction can be made between *potential* explanations and *actual* explanations. A potential explanation is a valid deductive argument the premises of which entail the conclusion. For an explanation to be an actual explanation, its singular premises must be true and its general premises must be well-confirmed lawful generalizations.

If the explanatory premises are true, a deductive argument provides conclusive evidence for the conclusion. When our premises are not so certain as to allow conclusive explanations, we have at least two alternatives: we can turn to probabilistic explanations (cf. [5, 25, 76]), or we can tentatively assume the premises are true and treat the explanation as *potentially refutable*.

The latter alternative is described by Popper [72]. In this approach, the premises of a deductive argument can be of two types: those that we accept as true, and those we treat as assumptions. These two types of premises taken together form a potential explanation of the deductively entailed conclusions. In the spirit of the scientific method, a potential explanation can be subjected to crucial experiments, and having survived the tests, the potential explanation becomes well-confirmed. Thus Popper writes:

> "A scientist, whether theorist or experimenter, puts forward statements, and tests them step by step. In the field of the empirical sciences, more particularly, he constructs hypotheses, or systems of theories, and tests them against experience by observation or experiment." [72, p. 27]
>
> "From a new idea, put up tentatively, and not yet justified in any way—an anticipation, a hypothesis, a theoretical system, or what you will—conclusions are drawn by means of logical deduction. These conclusions are then compared with one another and with relevant statements, so as to find what logical relations (equivalence, derivability, compatibility, incompatibility) exist between them." [72, p. 32]

This is the philosophy underlying the Theorist framework. Instead of viewing reasoning as deduction from our knowledge, reasoning may be better modelled by scientific theory formation. While the intuition underlying Theorist stems from the deductive nomological model of explanation, there is a notable difference. The Theorist framework does not require explanations to contain lawful general statements. Thus explanations in Theorist are usually only potential explanations. Because of this, it may be more appropriate to think of Theorist's explanations as "prescientific" (or common sense) theories.

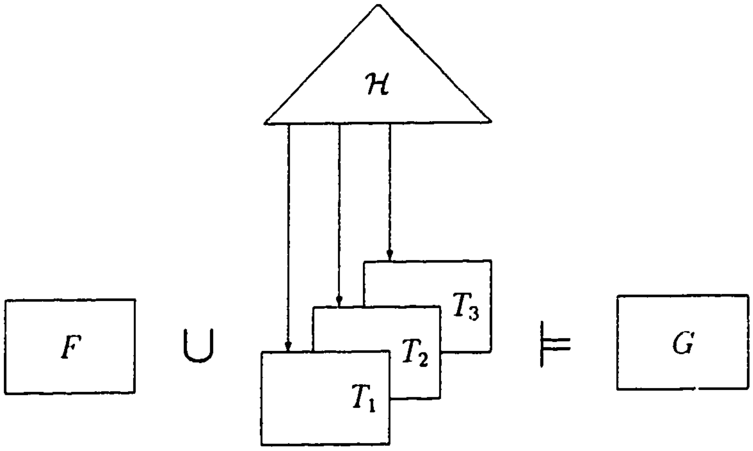
Even though Theorist's possible hypotheses are not usually lawful generalizations, Rescher's remarks about where laws come from furnish some insight into the often asked question, "where do Theorist's possible hypotheses come from." In the Theorist framework the answer is simply that possible hypotheses are supplied by the user and, in so doing, the user licenses Theorist to use the possible hypotheses in certain ways to form explanations.

2.2 Theorist Framework

A hypothetical reasoning framework should ultimately be concerned with observation, explanation, prediction, and comparing, testing and revising theories. The original Theorist framework [70], however, was only concerned with explaining observations. Later, Goodwin, Goebel and Gagné [17, 16] and Poole [69] proposed elaborations incorporating prediction; comparing theories was considered by Poole [62], Goodwin and Goebel [17, 15]; and theory testing and revision was discussed by Sattar, Goodwin and Goebel [77, 80, 79, 78]. In this section, the original version of Theorist is described, along with Poole's extension to include prediction and a distinction between conjectures, defaults, and conventions [70, 69, 66, 64].

The Theorist framework views reasoning as building theories that logically imply a set of observations. Statements of the underlying representation language, full first order clausal logic, are divided into two types: *facts* and *possible hypotheses*. Facts are statements that we accept as true and thus constitute ordinary assertions in a logical theory. Possible hypotheses, however, are given a different epistemological status than facts. They are used as a kind of schema for axioms—possible hypotheses can be viewed as the specification for the generation of logical theories which are extensions of the facts. Instances drawn from the set of possible hypotheses can be used to construct explanations for a set of observations.

A *theory* T, consisting of instances drawn from a set of *possible hypotheses* $\mathcal{H}$, is said to *explain* a set of *observations* G if the theory, together with the *facts* F, is consistent and logically implies the observations (see Figure 1). It is important to note that $F \cup T$ is an ordinary logical theory having standard Tarskian semantics.



T explains G if $T \subseteq \hat{\mathcal{H}}$[1] such that $F \cup T \models G$ and $F \cup T$ is consistent.

Figure 1: Theorist Framework

[1] H denotes the set of instances of elements of H w.r.t. the universe of discourse U.

The formation of instances of possible hypotheses depends on the underlying universe of discourse U. Most implementations of Theorist take U to be the Herbrand universe of $F \cup \mathcal{H} \cup G$ (as instantiation in Prolog implementations amounts to binding variables through unification), though the introduction of Skolem functions may enlarge U [11].

In the original Theorist framework [70], there is only one kind of possible hypothesis. If the user designates a sentence schema as a possible hypothesis, the user is licensing Theorist to use ground instances of that sentence schema in forming explanations, provided the ground sentence is consistent with the facts and everything else assumed. Poole [64, 68] later draws a distinction between *conjectures*, *defaults*, and *communication conventions*. In the next chapter, a means to distinguish between various kinds of hypotheses, each kind licensing a different use, is provided.

Explanation in Theorist depends on logical consequence and consistency (see Figure 1). For any logic as powerful as first order logic, each of the two conditions is, in general, only semi-decidable. Since we have good reason to want Theorist's underlying logic to be at least as powerful as first order logic [22], and since we are interested in actual computation, not just mathematical abstractions, we must inevitably be concerned with the inherent undecidability of such a powerful system. While undecidability is a problem in theory, it is not often encountered in many interesting problem domains —when it is, we must resort to heuristics. This should not be seen as a blemish on the methodology underlying Theorist. Rather, we should be surprised if scientific theory formation were decidable. In practice, we could, perhaps, submit our explanations to a partial consistency check. Passing this test can be viewed as confirming evidence for the potentially refutable explanation. This view of reasoning has a basis in the philosophy of science (cf. [72, 25, 76]).

In most implementations of Theorist, the internal representation language (the form in which expressions are represented in the underlying implementation language, e.g., Prolog) is full first order clausal logic; in this case, F, T, and G are sets of sentences, and $\mathcal{H}$ is a set of sentence schemas. Generally, the sentence schema in $\mathcal{H}$ are restricted to be schema for literals (i.e., $\hat{\mathcal{H}}$ is a set of ground literals). This in no way limits the expressive power of Theorist, as proved in [68], but it greatly simplifies Theorist's conceptual complexity and its implementation. Since $T \subseteq \hat{\mathcal{H}}$, theories are also sets of ground literals.

The sentences of the external representation language (the form in which expressions are provided by and presented to the user) of most Theorist implementations consist of predefined commands followed by their arguments. The commands to compile facts and hypotheses to their internal representation are as follows:

fact *Wff*. means that if *Wff* is a wff in first order predicate calculus[2] and $\{Clause_1, \ldots, Clause_n\}$ is a set of clauses corresponding to *Wff* then $\{Clause_1, \ldots, Clause_n\} \subseteq F$.

hypothesis *Hypothesis*. means that the literal $Hypothesis \in \mathcal{H}$.

hypothesis *Hypothesis*: *Wff*. is synonymous with

> **hypothesis** *Hypothesis*.
> **fact** $Clause_1 \leftarrow Hypothesis$.
> . . .
> **fact** $Clause_n \leftarrow Hypothesis$.

To illustrate the original Theorist framework, the well-known birds fly example is represented[3] in Example 2.1.

fact bird(tweety).
fact bird(X) ← penguin(X).
fact ¬fly(X) ← penguin(X).
hypothesis bf(X): fly(X) ← bird(X).

Example 2.1: Theorist Representation of Birds Fly

This encodes the knowledge that Tweety is a bird and all penguins are birds and penguins don't fly. As well, if something is a bird then it can be assumed to fly provided the assumption is consistent. From this, G_1 = {*fly(tweety)*} can be explained with the theory T_1 = {*bf(tweety)*} formed by instantiating the hypothesis *bf(X)* with X = *tweety*. Thus we can explain

[2] The extension allowing first order wffs in Theorist's external representation language is due to Ferguson [11].

[3] The syntax used in representations given in this chapter differs slightly from the implementation language (as ordinary keyboards lack some special symbols).

that Tweety flies because Tweety is a bird and birds can be assumed to fly. Should we later learn that

> **fact** penguin(tweety).

then G_1 can no longer be explained as T_1 is inconsistent with the facts. However $G_2 = \{\neg \mathit{fly(tweety)}\}$ can be explained with the theory $T_2 = \{\}$, i.e., Tweety does not fly follows deductively from the facts.

A typical dialogue with Theorist is given in Figure 2. The user interacts

```
fact bird(tweety).
fact bird(X) ← penguin(X).
fact ¬fly(X) ← penguin(X).
hypothesis bf(X): fly(X) ← bird(X).
explain fly(tweety).

Answer is: [fly(tweety)]
Theory is: [bf(tweety)]
No (more) answers.

fact penguin(tweety).
explain fly(tweety).

No (more) answers.

explain ¬fly(tweety).

Answer is: [¬fly(tweety)]
Theory is: []
No (more) answers.
```

Figure 2: Typical Theorist Dialogue

with the system by providing facts and hypotheses, and by posing queries (requests for explanations or predictions). In the figure, the user's input is shown in boldface. The system responds to queries with answers (which indicate variable bindings, if any) and theories. User interface features such as file input, knowledge base listing, tracing, etc. are provided in many of the various implementations.

In general, there may be multiple theories that explain the observations. An example of a set $\mathcal{E}_G$ of explanations of G for an arbitrary set of theories is illustrated in Figure 3. The set $\mathcal{U}$ contains the unacceptable theories, i.e., the inconsistent ones. In the figure, all the theories depicted are consistent theories except T_1. The ones in the set $\mathcal{E}_G$ entail G and are explanations of G. Note that T_2 is not an explanation of either G or $\neg G$. It may, however, be an explanation of some other goal.

Multiple theories in Theorist correspond to multiple minimal models in circumscription [46] and to multiple extensions in default logic [75]. It may be the case that some of these theories do not make intuitive sense. The problem of having multiple theories of which only a subset make intuitive sense is known as the multiple extension problem [20].

To illustrate, suppose in the birds fly example, it is possible that some penguins fly but that we can still assume penguins don't fly if it is consistent to do so. Also suppose Tweety is a penguin. This is represented in Example 2.2. From this, $G_1 = \{\mathit{fly(tweety)}\}$ can be explained with the the-

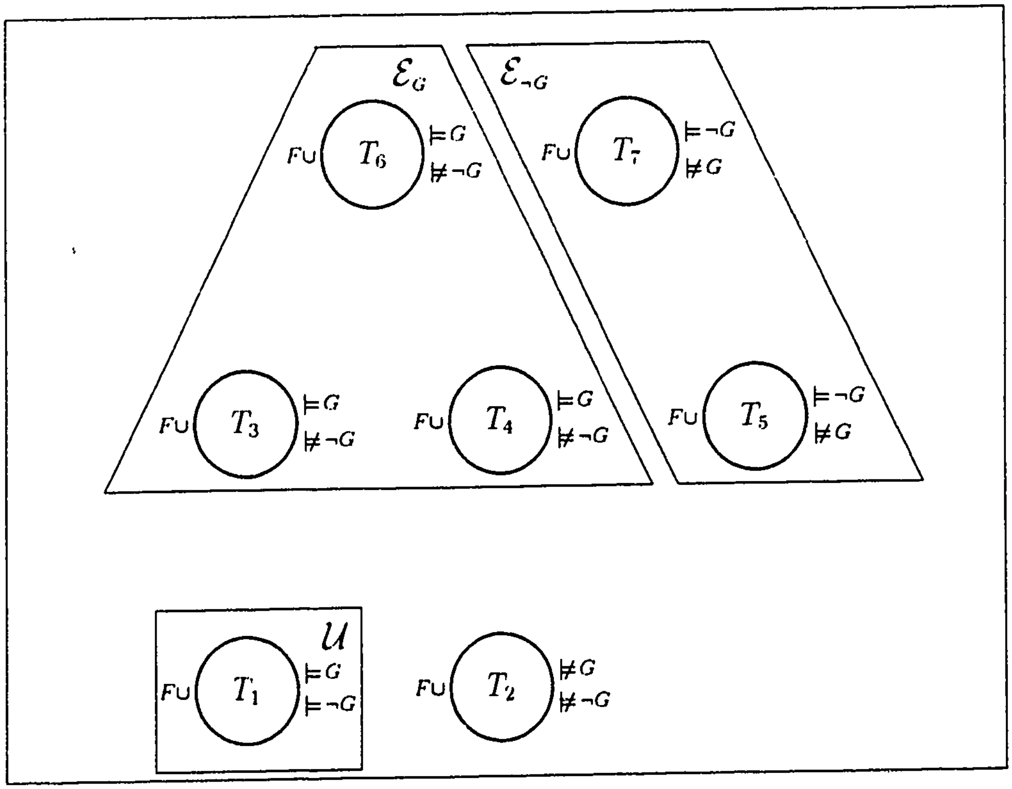


Figure 3: Explanations of a Goal

```
fact penguin(tweety).
fact bird(X) ← penguin(X).
hypothesis bf(X): fly(X) ← bird(X).
hypothesis pdf(X): ¬fly(X) ← penguin(X).
```

Example 2.2: Theorist Representation of Birds Fly and Penguins Don't ory T_1 = {*bf(tweety)*} but also, G_2 = {¬*fly(tweety)*} can be explained with the theory T_2= {*pdf(tweety)*}. But, intuitively, only the second theory makes sense. If Tweety did fly, it would be because he was a special kind of penguin (e.g., magical) or in special circumstances (e.g., on a plane) and not because Tweety is a bird.

This is not to say that Theorist is reasoning incorrectly in the above example. By designating *bf(X)* as a possible hypothesis, we licensed Theorist to use instances of this hypothesis in an explanation *whenever* it is consistent to do so. The problem is not with Theorist but that we gave Theorist too strong a license to use *bf(X)*. Yet the Theorist framework provides no mechanism to sanction a more restricted use of possible hypotheses. Consequently, the Theorist framework needs to be extended to solve the multiple extension problem.

Theorist can be extended to discriminated between multiple competing theories, in a way that is in accord with the philosophy of science, by incorporating theory preference and theory pruning knowledge. This meta-level knowledge can be used to prune theories that are irrelevant, or to order theories by utility or likelihood, etc. This extension of the basic framework is the subject of the next chapter.

Poole extended Theorist with another mode of reasoning: *prediction*. In prediction, we are interested, not in what we could assume to explain some goal or observation, but with whether all theories of a particular type have the goal as a consequence. To specify whether a theory is of the right type for prediction. Poole first distinguishes two types of possible hypotheses: *conjectures* and *defaults*. Conjectures are hypotheses having no particular justification. Defaults, on the other hand, are hypotheses that are expected to be true if they are consistent. Consequently, Poole defines prediction[4] as determining whether a goal is entailed by every theory that contains only instances of defaults and is *maximal* with respect to set inclusion. By virtue of the justification of defaults, such predicted goals are likewise expected to be true.

A goal G is *predicted* if G follows from every maximal theory (containing only defaults). The set of maximal theories $\mathcal{P}$ on which prediction is based is illustrated in Figure 4. The arcs in the figure represent the subset relation between theories which determines the maximal theories $\mathcal{P}$. $\mathcal{U}$ contains the inconsistent theory T_1. It also contains theories that contain conjectures (none are shown). (It will become clear in the next chapter that $\mathcal{P}$ and $\mathcal{U}$ correspond to theory preference and theory pruning criteria, respectively.) Since there is at least one maximal theory that does not entail G, it is not predicted. Similarly, $\neg G$ is not predicted either.

As a more concrete illustration, consider again Examples 2.1 and 2.2. If the hypotheses are considered defaults, then in the former example, we

[4] Other more skeptical and less skeptical definitions of *predict* are possible. See [64, 42].

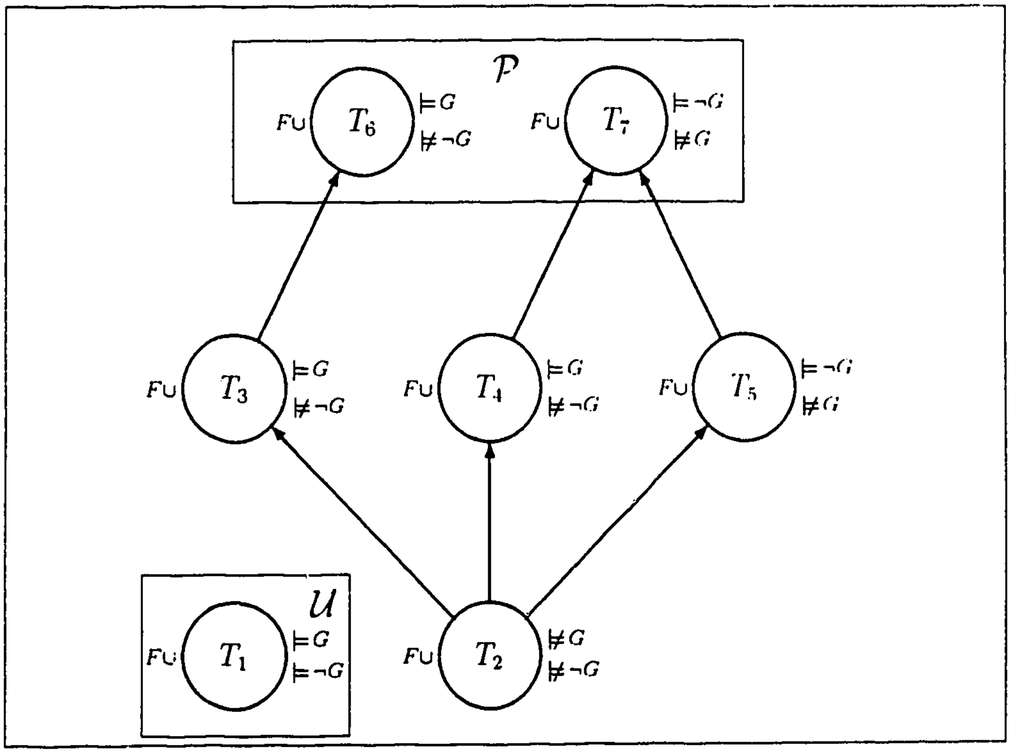

$\mathcal{P}$
$F \cup$ T_6 $\models G$ $\not\models \neg G$
$F \cup$ T_7 $\models \neg G$ $\not\models G$
$F \cup$ T_3 $\models G$ $\not\models \neg G$
$F \cup$ T_4 $\models G$ $\not\models \neg G$
$F \cup$ T_5 $\models \neg G$ $\not\models G$
$\mathcal{U}$
$F \cup$ T_1 $\models G$ $\models \neg G$
$F \cup$ T_2 $\not\models G$ $\not\models \neg G$

Figure 4: Prediction from all Maximal Theories

predict $G_1 = \{\mathit{fly(tweety)}\}$ because $T_1 = \{\mathit{bf(tweety)}\}$ is the unique maximal theory and it entails G_1. In the latter example, $T_1 = \{\mathit{bf(tweety)}\}$ and $T_2 = \{\mathit{pdf(tweety)}\}$ are both maximal so neither $G_1 = \{\mathit{fly(tweety)}\}$ nor $G_2 = \{\neg \mathit{fly(tweety)}\}$ is predicted.

In the next section, the implementation of a Theorist Interpreter is discussed.

2.3 Theorist Interpreter

Many versions of Theorist have been implemented in Prolog, e.g., Ferguson's C-Theorist [11], Poole and Goodwin's Q-Theorist compiler [71, 67], and Poole, Goebel, and Aleliunas's Theorist interpreter [70]. These implementations are essentially based on a first order logic theorem prover extended to manage assumptions and provide user interface features.

In Poole's approach, there are only two kinds of hypotheses: conjectures and defaults. Let $\mathcal{C}$ be the subset of $\mathcal{H}$ interpreted as conjectures and Δ be the subset of $\mathcal{H}$ interpreted as defaults. A few definitions and theorems are necessary to set the stage for the definition of procedures for explanation and prediction.

Definition 1 (SCENARIO (POOLE [64, P. 12]))

A `scenario` of $F \cup \mathcal{C} \cup \Delta$ is a set $F \cup T$ where T is a set of ground instances of elements of $\mathcal{C} \cup \Delta$ such that $F \cup T$ is consistent.

Scenarios are viewed as being possibly true in the domain of interest.

Definition 2 (EXPLANATION FROM $F \cup \mathcal{C} \cup \Delta$ (POOLE [64, P. 13]))
If g is a closed formula then an **explanation** of g from $F \cup \mathcal{C} \cup \Delta$ is a scenario of $F \cup \mathcal{C} \cup \Delta$ which implies g.

This definition allows defaults to be used in explanations as long as they are consistent with the facts and other assumptions. Poole gives the following theorems to provide a basis for implementation of the above definition of explanation.

Theorem 3 (EXPLANATION PROOFS (POOLE [64, P. 34]))
If F is consistent and $T = \{t_1, \ldots, t_n\}$ is a set of ground instances of elements of $\mathcal{C} \cup \Delta$ then $F \cup T$ is an explanation of g iff there is a ground proof of g from $F \cup T$ such that $F \cup \{t_1, \ldots, t_{i-1}\} \not\vdash \neg t_i$ for all $i = 1 \ldots n$.

From this Poole [64, p. 34] derives the following "algorithm[5]" to explain g from $F \cup \mathcal{C} \cup \Delta$:

1. try to prove g from $F \cup \mathcal{C} \cup \Delta$ and make $T = \{t_1, \ldots, t_n\}$ the set of instances of elements of $\mathcal{C} \cup \Delta$ used in the proof.

2. ground T so we have created a ground proof of g from $F \cup T$.

[5] The "algorithm" is, of course, not even semi-decidable in general.

3. for each $t_i \in T$, try to prove $\neg t_i$ from $F \cup \{t_1, \ldots, t_{i-1}\}$. If all such proofs fail, $F \cup T$ is an explanation of g from $F \cup \mathcal{C} \cup \Delta$.

An important feature of this procedure is that step 3 allows for *incremental consistency checking*, i.e., consistency can bc checked as each new hypothesis instance is added to the current theory. Figure 5 gives a Quintus Prolog implementation of this procedure for explanation. The predicate `explain` is true when the goal `Goal` is explained by the facts and the consistent set of hypothesis instances `Theory`. The predicate `prove` is a first order logic theorem prover based on Loveland's MESON proof procedure [43] and augmented with the ability to introduce consistent instances of hypotheses. The first clause of `prove` checks the list of ancestors goals for the negation of the current goal, i.e., it checks for proof by contradiction. The second clause attempts to prove the goal from the facts. The third clause tries to use a previously assumed hypothesis instance and the final clause allows additional consistent assumptions. Step 2 of the procedure is not included in the given implementation (cf. [70]) which is only correct when there are no Skolem functions and there are no free variables in the hypotheses when they are tested for consistency—this problem has been addressed by Ferguson [11]. Figure 6 shows the implementation of the test for inconsistency.

To implement prediction from defaults, Poole defines maximal scenarios (which have corresponding maximal theories).

```
explain(Goal,Theory) :-
    prove(Goal,[],[],Theory).


prove(Goal,Ancestors,InTheory,InTheory) :-
    negate(Goal,NotGoal),
    member(NotGoal,Ancestors).

prove(Goal,Ancestors,InTheory,OutTheory) :-
    fact(Goal,IfBody),
    proveAll(IfBody,[Goal|Ancestors],InTheory,OutTheory).

prove(Goal,Ancestors,InTheory,InTheory) :-
    hypothesis(Goal),
    member(PreviousHypothesis,InTheory),
    Goal == PreviousHypothesis.

prove(Goal,Ancestors,InTheory,[Goal|InTheory]) :-
    hypothesis(Goal),
    not( member(PreviousHypothesis,InTheory),
         Goal == PreviousHypothesis ),
    not( inconsistent([Goal|InTheory]) ).


proveAll([],Ancestors,InTheory,InTheory).

proveAll([Goal|Goals],Ancestors,InTheory,OutTheory) :-
    InTheory == OutTheory,
    prove(Goal,Ancestors,InTheory,InTheory),
    proveAll(Goals,Ancestors,InTheory,InTheory).

proveAll([Goal|Goals],Ancestors,InTheory,OutTheory) :-
    not( InTheory == OutTheory ),
    prove(Goal,Ancestors,InTheory,IntermedTheory),
    proveAll(Goals,Ancestors,IntermedTheory,OutTheory).
```

Figure 5: Explain in Prolog

```
inconsistent([Goal|Theory]) :-
    negate(Goal,NotGoal),
    not( prove(NotGoal,[],InTheory,InTheory) ).
```

Figure 6: Inconsistent in Prolog

Definition 4 (EXTENSIONS (POOLE [64, P. 13]))

The set of logical consequences of a maximal (with respect to set inclusion) scenario S of $F \cup \Delta$, written $Th(S)$, is an **extension** of $F \cup \Delta$.

Poole [68] proves the correspondence between this definition of extension and that of Reiter [75] when $\delta \in \Delta$ corresponds to Reiter's default :Mδ/δ.

Theorem 5 (EXPLANATIONS AND EXTENSIONS (POOLE [64, P. 13]))

There is an explanation of g from $F \cup \Delta$ iff there is some extension of $F \cup \Delta$ that contains g.

Definition 6 (PREDICTION FROM $F \cup \Delta$ (POOLE [64, P. 13]))

We **predict** g based on $F \cup \Delta$ if g is in every extension of $F \cup \Delta$. (Note that no conjectures are involved in (strict) prediction.)

An procedure for prediction can be derived from the following theorem.

Theorem 7 (PREDICTION PROOFS (POOLE [64, P. 34]))

The following are equivalent:

1. g is in every extension of $F \cup \Delta$.

2. every scenario S of $F \cup \Delta$ is an explanation of g.

3. there is a set $\mathcal{E}_g$ of (finite) explanations of g such that every scenario S of $F \cup \Delta$ is consistent with some $E \in \mathcal{E}_g$.

4. there is some set T of instances of Δ such that $F \cup T \vdash g$ such that if $t_i \in T$ and $\neg t_i$ is explainable by T_i, then g is in every extension of $F \cup T_i \cup \Delta$.

From point 4 of this theorem, Poole [64, p. 36] suggests the following procedure for proving that g is in every extension of $F \cup \Delta$, i.e., that g is predicted from $F \cup \Delta$:

1. try to prove g from $F \cup \Delta$ and make $T = \{t_1, \ldots, t_n\}$ the set of instances of elements of Δ used in the proof.

2. ground T so we have created a ground proof of g from $F \cup T$.

3. for each $t_i \in T$, try to explain $\neg t_i$ from $F \cup \{t_1, \ldots, t_{i-1}\} \cup \Delta$. If there is an explanation using no assumptions then T is inconsistent; otherwise for each T_i explaining $\neg t_i$, try to prove g is in all extensions of $F \cup T_i \cup \Delta$.

Figure 7 gives a Quintus Prolog implementation of this procedure for prediction. The predicate `predict` is true when `Goal` can be derived from every maximal theory (i.e., it is in all extensions). The predicate `inAllExtensions` is true if *Goal* is in all extensions of $F \cup S \cup \Delta$. Finally, the predicate `counterExplanation` is true if there is an explanation `NewS` of the negation of some assumption Ti in T that includes the assumptions in S and if Goal is not in every extension of the explanation NewS. As step 2 is not included in this implementation, the same proviso given for *explain* holds for *predict*. As well, this implementation assumes all hypotheses are defaults.

```
predict(Goal) :-
    inAllExtensions(Goal,[]).


inAllExtensions(Goal,S) :-
    prove(Goal,[],S,T),
    not( counterExplanation(Goal,S,T) ).

counterExplanation(Goal,S,T) :-
    member(Ti,T),
    negate(Ti,NotTi),
    prove(NotTi,[],S,NewS),
    not( inAllExtensions(Goal,NewS) ).
```

Figure 7: Predict in Prolog

The Theorist framework described above is simple, yet powerful. It has been applied to several domains such as inheritance reasoning (e.g., [62]), temporal reasoning (e.g., [15, 14]), analogical reasoning (e.g., [27, 86, 13]), diagnosis (e.g., [63]), and many others. In the next chapter, our extension to the Theorist framework is presented.

Chapter 3

Meta-Theorist

This chapter develops a hypothetical reasoning framework, called Meta-Theorist, which is our extension of Theorist. Representing and reasoning in Meta-Theorist involves specifying what is known about the domain; what can be assumed about the domain; under which conditions assumptions are justified; and what can be inferred from the above. Several kinds of hypotheses are described in this chapter—each can be considered as an assumption plus conditions for its applicability. Meta-Theorist incorporates meta-level theory pruning and theory preference knowledge that can be used to eliminate unacceptable or irrelevant theories, or to order theories by utility or likelihood, etc. This extension to Theorist subsumes many previous extensions to Theorist in a conceptually efficient way. Meta-Theorist is a general tool for hypothetical reasoning, but in developing it we have in mind its use in specifying and experimenting with defaults (see Chapter 6).

3.1 Meta-Theorist Framework

3.1.1 The Meta-Level

The basic Theorist framework is extended with a set of meta-level sentences $\mathcal{M}$ which allows (among other things) the specification of theory preference and theory pruning criteria. As with the object level sentences, the underlying representation language is full first order clausal logic. The external representation language provides a command to compile meta-level wffs to their internal form.[1]

> **meta fact** *Wff*. means that if *Wff* is a wff in first order predicate calculus and $\{Clause_1, \ldots, Clause_n\}$ is a set of clauses corresponding to *Wff* then $\{Clause_1, \ldots, Clause_n\} \subseteq \mathcal{M}$.

Sentences and sets of sentences are objects in the semantic domain of the meta-level and so F, $\hat{\mathcal{H}}$, theories, etc. may be used as arguments in meta-level predicates. The commands to compile facts and hypotheses are now reflected (cf. section 2.2) in the meta-level as follows:

> **fact** *Wff*. is synonymous with
>
> **meta fact** fact($Clause_1$).
>
> ...

[1] The set of meta-level commands also includes **meta hypothesis**, **meta explain**, **meta predict**, etc. The extension also provides a meta-meta-level and beyond. Discussion of these features is omitted as they are not of prime concern to this dissertation.

meta fact fact($Clause_n$).

if *Wff* is a wff in first order predicate calculus and $\{Clause_1, \ldots, Clause_n\}$ is a set of clauses corresponding to *Wff*.

hypothesis *Hypothesis*. is synonymous with

meta fact hypothesis(*Hypothesis*).

hypothesis *Hypothesis*: *Wff*. is synonymous with

hypothesis *Hypothesis*.
fact $Clause_1 \leftarrow Hypothesis$.
...
fact $Clause_n \leftarrow Hypothesis$.

Two meta-level predicates of special interest are: *prefer(T1,T2)* and *prune(T)*. The procedures for explanation and prediction (i.e., the meta-predicates *explain(G,T)*, *preferred_explanation(G,T)*, and *predict(G)*), use these two predicates to determine theory preference and theory pruning. The theory pruning and theory preference meta-predicates considered below are defined independently of the goal and independently of the kind of inference. One might envision certain goal-dependent or inference-dependent pruning and preference criteria. Such criteria can be expressed by a straightforward extension.

The user's interaction with Meta-Theorist is essentially the same as with Theorist. The main difference is that the user can specify meta-facts, meta-hypotheses, etc. and that these meta-level statements can be used to specify theory pruning and preference criteria. Additionally, we might have a system designer/knowledge engineer who provides the user with a pre-packaged set of pruning/preference criteria for particular domains. For instance in Chapter 6, we envision defaults to be a kind of hypothesis with an associated pruning criteria provided by the system designer.

3.1.2 Theory Pruning

The conditions under which a theory is deemed unacceptable (e.g., inconsistent, irrelevant, etc.) define the prune predicate. These conditions may be specified via:

meta fact prune(T) ← *Wff.*

If $\mathcal{M} \models$ *prune(T)* then the theory T is in the set of ***unacceptable theories*** $\mathcal{U}$. By $\mathcal{U}(T)$ it is meant that T is unacceptable according to the pruning criteria determining $\mathcal{U}$, i.e., $T \in \mathcal{U}$. Throughout this chapter, the following completion of the set of pruning criteria is assumed.[2]

meta fact ¬prune(T) ← naf(prune(T)).

The set of ***possible theories*** Π is defined as the power set of $\hat{\mathcal{H}}$ excluding elements of $\mathcal{U}$.[3] By $\Pi(T)$, we mean $T \in \Pi$. We are generally only interested in possible theories (e.g., theory preference in section 3.1.3 is only defined for possible theories).

meta fact possible(T) ← theory(T) & ¬prune(T).

[2] The meta-meta-predicate *naf* means negation-as-failure.

[3] The meta-predicate *theory(T)* means $T \subseteq \hat{\mathcal{H}}$.

Inconsistent theories are considered impossible theories (though in view of the semi-decidability of consistency checking, we might want to treat theories that have not yet been shown inconsistent as possible theories — we do not intend to pursue this here). Throughout the rest of the chapter, representations are implicitly assumed to include:[4]

meta fact prune(T) ← inconsistent(T).
meta fact inconsistent(T) ↔
    ∃H (member(H,T) & prove(¬H,T)).
meta fact consistent(T) ↔ ¬inconsistent(T).

Recall that in the modified birds fly example (Example 2.2), there was a counterintuitive explanation T_1 (see section 2.2) that explained Tweety flies even though Tweety is a penguin. We could eliminate T_1 by providing a pruning criterion based on specificity [62, 30], that is, the hypothesis that penguins don't fly is more specific knowledge than the hypothesis that birds fly, so whenever the former hypothesis is applicable (i.e., when the bird in question is known to be a penguin), the latter hypothesis is not applicable. This might be naively represented as follows.[5]

meta fact prune(T) ← minus(T,[bf(X)],T2) &
    minus(T1,[pdf(X)],T2) & consistent(T1) &
    union(T,[pdf(X)],T3) & ¬consistent(T3).

[4] The meta-predicate *member(H,T)* means H∈T, and *prove(P,T)* is true whenever F∪T⊨P.

[5] The meta-predicate *minus(T,[H],T1)* means T-{H}≡T1, and *union(T,[H],T1)* means T∪{H}≡T1.

That is, the theory T should be pruned if it contains a *bf* assumption but a corresponding *pdf* assumption cannot be consistently added while the *bf* assumption could be consistently replaced by a corresponding *pdf* assumption. The above criterion is not being proposed as a correct formalization of specificity. There is much debate on how this intuitive concept should be formalized [8]. The above is just a naive formalization to illustrate the use of pruning. The issue of specificity arises again in connection with defaults in the next three chapters.

3.1.3 Theory Preference

Various kinds of theory preference knowledge can be used to rank theories (e.g., simplicity, predictive power, likelihood, etc.). Theory preference is described as a partial order, *prefer*, on Π, the set of possible theories (cf. the partial order $\geq^p$ in [17, 14]). When we say one theory is preferred to another, we mean the former is better or equal to the latter. Theory preference (i.e., when $T1$ is preferred over $T2$) is asserted in the following manner:

meta fact prefer(T1,T2) ← *Wff*.

To be a partial order, *prefer*, must be *reflexive*, *transitive*, and *antisymmetric*, i.e., it satisfies the following axioms:

meta fact prefer(T,T).
meta fact prefer(T1,T3) ← prefer(T1,T2) & prefer(T2,T3).
meta fact equipreferable(T1,T2) ←
    prefer(T1,T2) & prefer(T2,T1).

Here the meta-predicate *equipreferable(T1, T2)* means $T1$ and $T2$ are equally preferred. Thus, it follows that

> **meta fact** prefer(T1,T2) ← equipreferable(T1,T2).

Some theories in the partial ordering may be *incomparable*, i.e., neither theory is preferred and the theories are also not equipreferable.

> **meta fact** incomparable(T1,T2) ←
> ¬prefer(T1,T2) & ¬prefer(T2,T1).

Throughout this chapter, the following completion of the set of preference criteria is assumed.

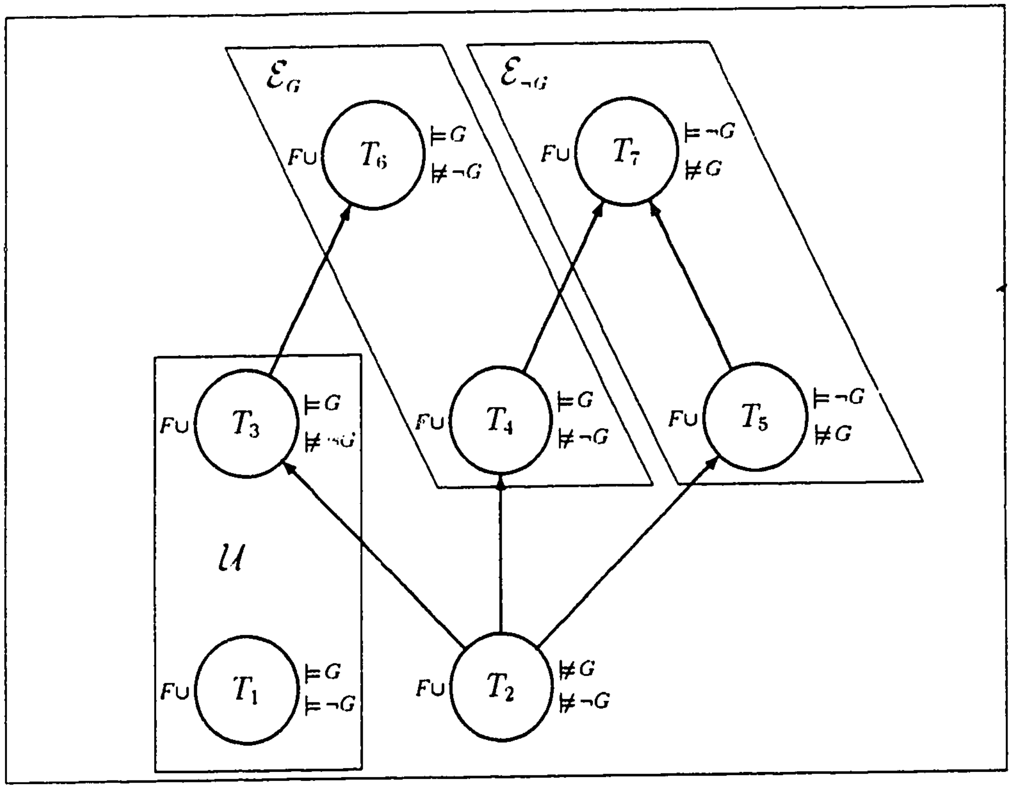
> **meta fact** ¬prefer(T1,T2) ← naf(prefer(T1,T2)).

3.1.4 Explanation and Prediction

The set of *viable* explanations $\mathcal{E}_G$ of a set of observations G are the elements of the set of possible theories Π which entail G given the facts F (See Figure 8 here the arcs represent preferences.). In Meta-Theorist, the set of possible theories is restricted by the pruning criteria, as the inclusion of T_3 in $\mathcal{U}$ suggests. Consequently, the definition of *explain* must be revised to take the theory pruning criteria into account. A theory T is said *explain* a set of observations G if $F \cup T \models G$ and $T \in \Pi$ (so $\mathcal{M} \not\models prune(T)$).

> **meta fact** explain(G,T) ← possible(T) & prove(G,T).

The definition of *explain* is meant as a specification—obviously there are more efficient ways to implement it.



$\mathcal{E}_G$
$\mathcal{E}_{\neg G}$
$F\cup$ T_6 $\models G$ $\not\models \neg G$
$F\cup$ T_7 $\models \neg G$ $\not\models G$
$F\cup$ T_3 $\models G$ $\not\models \neg G$
$F\cup$ T_4 $\models G$ $\not\models \neg G$
$F\cup$ T_5 $\models \neg G$ $\not\models G$
$\mathcal{U}$
$F\cup$ T_1 $\models G$ $\models \neg G$
$F\cup$ T_2 $\not\models G$ $\not\models \neg G$

Figure 8: Explanations of a Goal in Meta-Theorist

Often we are interested in finding preferred explanations. The set $\mathcal{P}_G$ of preferred theories explaining G are the maximal elements (w.r.t. prefer) of the subset of Π that explain G (Figure 9).

meta fact preferred_explanation(G,T1) ← explain(G,T1) & ∀T2 (equipreferable(T1,T2) ← prefer(T2,T1) & explain(G,T2)).

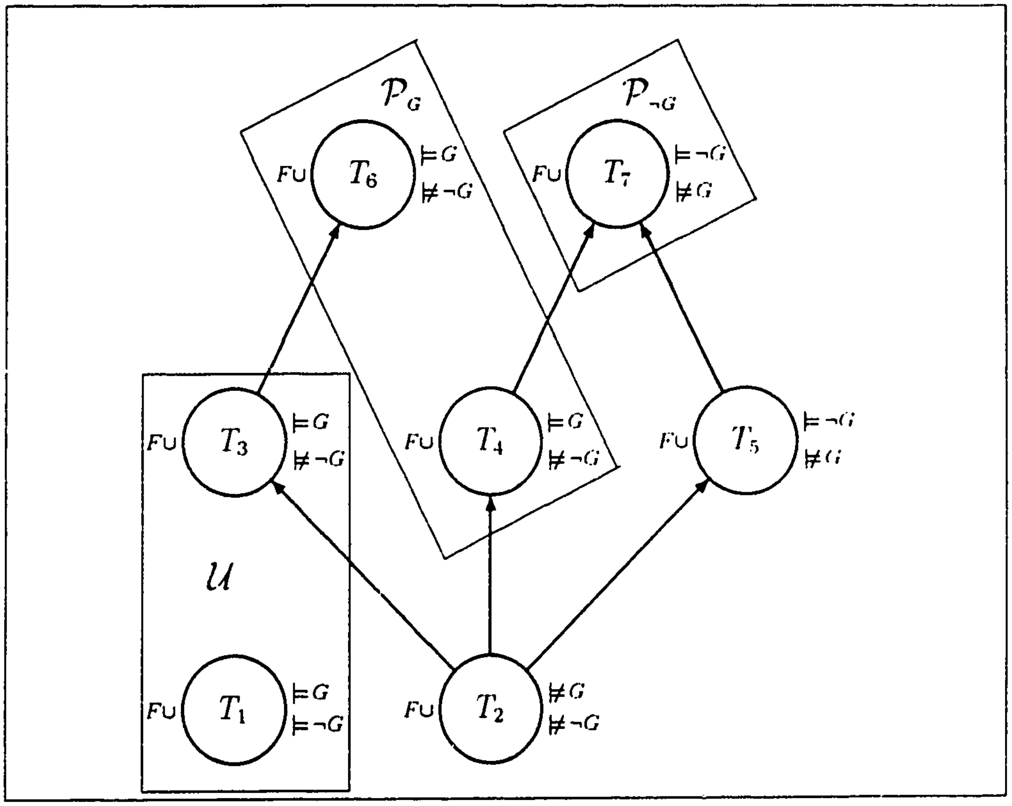


Figure 9: Preferred Explanations of a Goal in Meta-Theorist

In prediction, we generally want to base our conclusions on the most preferred theories. The most preferred or *maximal* elements of the partial order are those for which there is no strictly better theory. A partial ordering may have multiple maximal elements but if it has a unique maximal element then that element is the *maximum* element.

```
meta fact maximal(T1) ←
    ∀T2 (equipreferable(T1,T2) ← prefer(T2,T1)).
meta fact maximum(T1) ← ∀T2 ((T1 = T2) ← prefer(T2,T1)).
```

We *predict* a goal *G* if it follows from every possible maximal theory.

```
meta fact predict(G) ←
    ∀T (prove(G,T) ← possible(T) & maximal(T)).
```

The set of preferred theories $\mathcal{P}$ on which prediction is based is shown in Figure 10. A comparison between Figure 9 and 10 reveals the essential difference between preferred explanation and prediction. Preferred explanation considers the maximal theories w.r.t. the partial order on the subset of Π which explain the goal. Prediction considers the maximal theories w.r.t. the partial order on all of Π regardless of whether the goal is explained.

The Meta-Theorist framework is illustrated in Figure 11.

3.2 Kinds of Hypotheses

The semantics of hypotheses is determined in part by the first order sentences they are schemas for and in part by theory preference and pruning. We can, in effect, define different kinds of hypotheses, licensing them to be used in different ways, by making the theory preference and pruning criteria dependent on the kind of hypotheses contained in the theories.

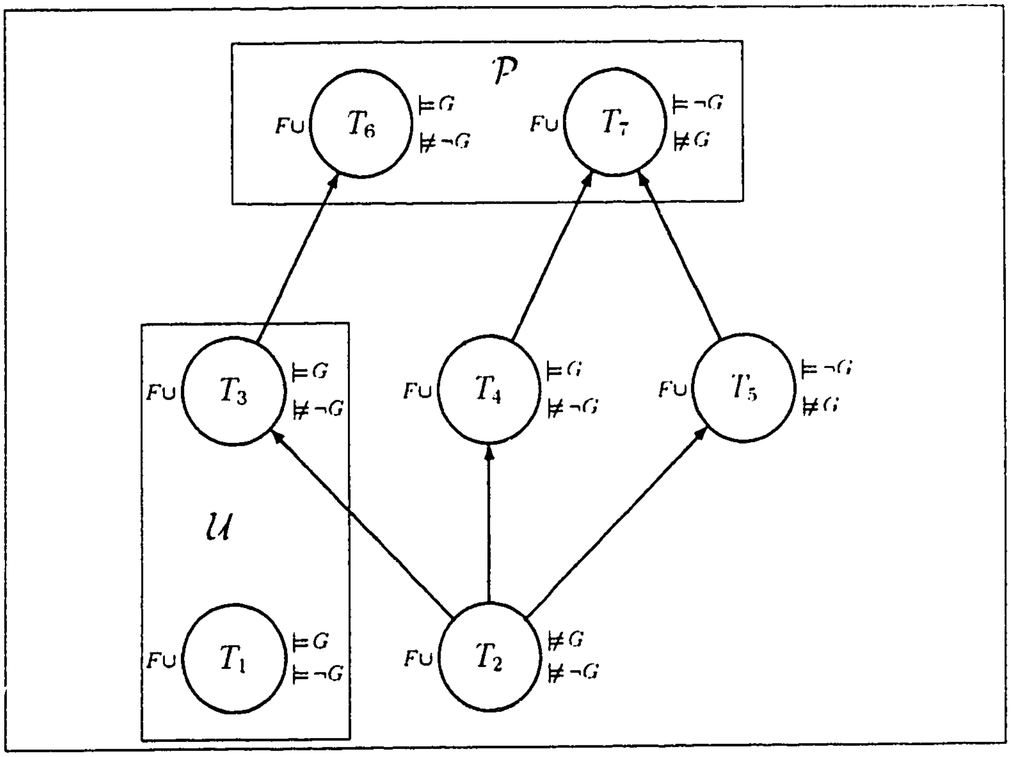
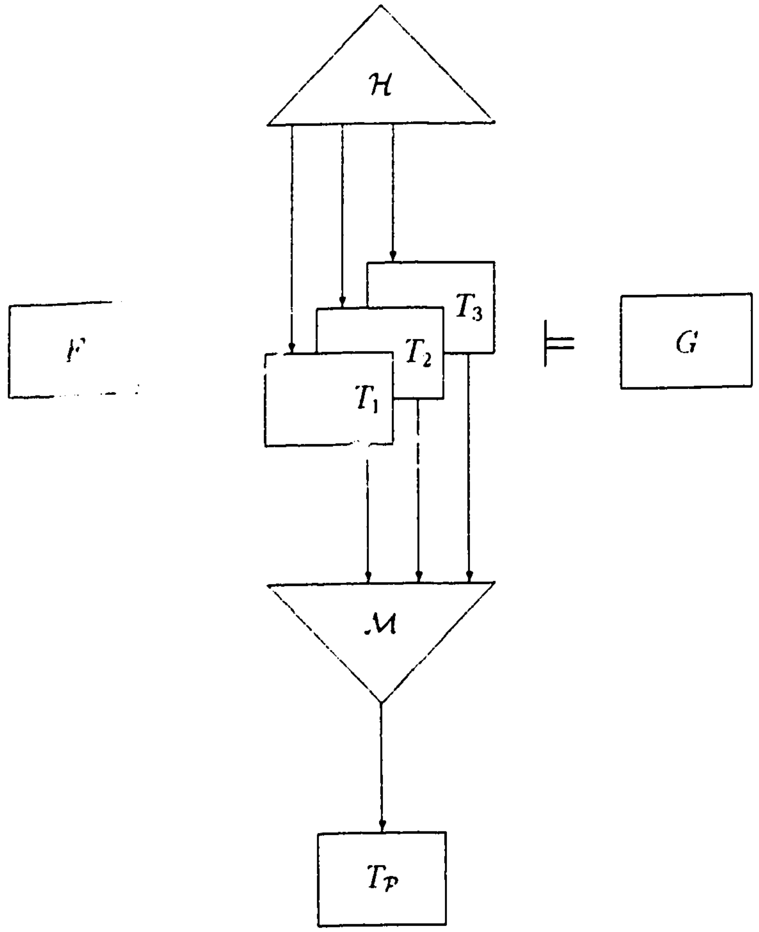

$\mathcal{P}$
$F\cup$ T_6 $\models G$ $\not\models \neg G$
$F\cup$ T_7 $\models \neg G$ $\not\models G$
$F\cup$ T_3 $\models G$ $\not\models \neg G$
$F\cup$ T_4 $\models G$ $\not\models \neg G$
$F\cup$ T_5 $\models \neg G$ $\not\models G$
$\mathcal{U}$
$F\cup$ T_1 $\models G$ $\models \neg G$
$F\cup$ T_2 $\not\models G$ $\not\models \neg G$

Figure 10: Preferred Theories Predicting a Goal in Meta-Theorist



T explains G if $T \in \Pi$ such that. $F \cup T \models G$
T is a preferred explanation of G if T explains G and $T \in \mathcal{P}_G$
G is predicted if $\forall T \in \mathcal{P}$. $F \cup T \models G$.

Figure 11: Meta-Theorist Framework

For our purposes, it is be convenient to partition the set of possible hypotheses into several classes. The most prevalent kind of hypothesis in the nonmonotonic reasoning literature is the default [75, 70]. The previous chapter introduced defaults in connection with prediction. Under the usual interpretation, defaults are assumed true in the absence of evidence to the contrary.

This interpretation of defaults is ambiguous because it is not clear what constitutes contrary evidence—particularly troublesome is the status of mutually inconsistent defaults—the resulting "clash of intuitions" led to a distinction between *credulous* defaults and *skeptical* defaults [83]. The difference is that conflicting credulous defaults lead to alternative theories while conflicting skeptical defaults are mutually blocking.

Another ambiguity is the status of default conclusions. Usually, they are taken to be only plausible conclusions. This leads to a couple of problems. First, the interpretation of the conclusions of Meta-Theorist change from "true if the assumptions are true" to "plausible if the assumptions are true." Second, is the problem of *chaining*, i.e., the plausible conclusion of one default may or may not be considered sufficient to "enable" another default.

A second interpretation of default conclusions is that they are true rather than merely plausible. This interpretation corresponds to communication conventions [47, 64]. For example, we might establish the convention that if I tell you something is a bird then I am in effect telling you it flies unless I tell you otherwise. So when I tell you Tweety is a bird, you can take it as a fact that Tweety is not a penguin because if Tweety were a penguin I should have told you so. While plausible defaults may be credulous or skeptical with regard to conflicts, communication conventions should not permit conflicts otherwise the convention (if one may call it that) is ambiguous [64].

Let us now consider how these various kinds of hypotheses might be specified in Meta-Theorist. The following specifications are not claimed to be correct. They are only intended to illustrate how various kinds of hypotheses might be specified.

Let Δ_{cd}, Δ_{sd}, and Δ_{cc} be subsets of $\mathcal{H}$ containing hypotheses that are to be interpreted as credulous and skeptical defaults respectively. The commands to specify these defaults are:[6]

credulous *Hypothesis.* is synonymous with
> **hypothesis** *Hypothesis.*
> **meta fact** credulous(*Hypothesis*).

credulous *Hypothesis*: *Wff.* is synonymous with
> **hypothesis** *Hypothesis*: *Wff.*
> **meta fact** credulous(*Hypothesis*).

skeptical *Hypothesis.* is synonymous with
> **hypothesis** *Hypothesis.*
> **meta fact** skeptical(*Hypothesis*).

skeptical *Hypothesis*: *Wff.* is synonymous with
> **hypothesis** *Hypothesis*: *Wff.*

[6] The meta-predicate *credulous(H)* means H$\in \Delta_{cd}$, *skeptical(H)* means H$\in \Delta_{sd}$ *convention(H)* means H$\in \Delta_{cc}$.

meta fact skeptical(*Hypothesis*).

convention *Hypothesis.* is synonymous with

hypothesis *Hypothesis.*
meta fact convention(*Hypothesis*).

convention *Hypothesis*: *Wff.* is synonymous with

hypothesis *Hypothesis*: *Wff.*
meta fact convention(*Hypothesis*).

Different semantics can by given to these by defining pruning criteria that treat elements of Δ_{cd}, Δ_{sd}, Δ_{cc} differently than other elements of $\mathcal{H}$. The following pruning criterion corresponds to a credulous interpretation of defaults.

meta fact prune(T) ← credulous(H) & ¬member(H,T) &
union(T,[H],T1) & consistent(T1).

This prunes any theory which can be consistently extended by an instance of a credulous default—this corresponds to the defaults in the previous chapter.

We could re-represent the birds fly example by treating the hypotheses as credulous defaults (Example 3.1). Consider the following theory:

fact bird(tweety).
fact bird(X) ← penguin(X).
credulous bf(X): fly(X) ← bird(X).
credulous pdf(X): ¬fly(X) ← penguin(X).

Example 3.1: Birds Fly using Credulous Defaults

T_1 = {bf(tweety),pdf(tweety)}.

This theory is the unique maximal[7] theory since it is the only possible theory (unless the goal enlarges the universe of discourse by mentioning another individual). From this we predict (cf. section 3.1.4) Tweety flies and is not a penguin.

Suppose we also know that:

fact penguin(opus).

In this case, the theory

$$T_2 = \{\mathrm{bf(opus).pdf(opus).bf(tweety).pdf(tweety)}\}$$

is not maximal since it is inconsistent. The conflicting credulous defaults *pdf(opus)* and *bf(opus)* give rise to the alternative maximal theories:

$$T_3 = \{\mathrm{pdf(opus).bf(tweety).pdf(tweety)}\}.$$
$$T_4 = \{\mathrm{bf(opus).bf(tweety).pdf(tweety)}\}.$$

If we use a specificity criterion similar to that of section 3.1.2 then theory T_3 would be the unique maximal theory from which we predict Opus does not fly.

For skeptical defaults, the following pruning criterion applies:

```
meta fact prune(T) ← skeptical(H) & ¬member(H,T) &
    & max_nonskeptical_subset(T,T1) & union(T1,[H],T2) &
    consistent(T2) & ¬blocked(H,T1).
```

[7] In the absence of any preference criteria, all possible theories are maximal.

This prunes any theory which can be extended by a consistent instance of a non-blocked skeptical default. Here *max_nonskeptical_subset(T,T1)* means *T*1 is the largest subset of *T* not containing any skeptical defaults this makes "skepticism" relative to other kinds of assumptions an even more skeptical approach would be to check consistency of *H* relative to the facts alone. A skeptical default is not blocked if it is not contradicted by any consistent set of skeptical defaults.

> **meta fact** ¬blocked(H,T1) ← ∀T2 ((∀H2 (member(H2,T2) →
> skeptical(H2)) & union(T1,T2,T3) & consistent(T3) &
> union(T3,[H],T4)) → consistent(T4)).

We could also re-represent the birds fly example by treating the hypotheses as skeptical defaults (Example 3.2). The difference from the previous example is that now if we know

```
fact bird(tweety).
fact bird(X) ← penguin(X).
skeptical bf(X): fly(X) ← bird(X).
skeptical pdf(X): ¬fly(X) ← penguin(X).
```

Example 3.2: Birds Fly using Skeptical Defaults

> **fact** penguin(opus)

the skeptical default instances *bf(opus)* and *pdf(opus)* are mutually blocking thus leaving

$$T_1 = \{\mathrm{bf(tweety)}, \mathrm{pdf(tweety)}\}$$

as the unique maximal theory. We could incorporate specificity by modifying the definition of blocking for skeptical defaults so that less specific defaults do not block more specific ones.

```
meta fact ¬blocked(H.T1) ← ∀T2 ((∀H2 (member(H2.T2) →
    (skeptical(H2) & ¬ [illegible]_specific(H.H2)) & union(T1.T2.T3)
    & consistent(T3) & union(T3.[H].T4)) → consistent(T4)).
```

This would give the unique maximal theory

$$T_3 = \{\text{pdf(opus).bf(tweety).pdf(tweety)}\}$$

from which we predict Opus does not fly.

The following pruning criterion corresponds to one possible interpretation of communication conventions:

```
meta fact prune(T) ← convention(H) & ¬member(H.T) &
    consistent(H).
```

This prunes any theory which can be extended by an instance of a consistent communication convention. Note that if there are conflicting communication conventions that are each consistent with the facts then every theory is pruned indicating that our communication conventions are ambiguous. A more elaborate interpretation of communication conventions might consider conflicting chains of conventions similar to blocking skeptical defaults. Another elaboration would consider the interaction of communication conventions with conjectures which are described later in this section. As it stands above, conventions are relative to the facts; but we may want them relative to the facts plus any conjectures we have assumed (see the discussion in section 3.3).

The birds fly example with hypotheses treated as communication conventions is given in Example 3.3. Here the unique maximal theory is:

```
fact bird(tweety).
fact bird(X) ← penguin(X).
convention bf(X): fly(X) ← bird(X).
convention pdf(X): ¬fly(X) ← penguin(X).
```

Example 3.3: Birds Fly using Communication Conventions

$$T_1 = \{\text{bf(tweety)}, \text{pdf(tweety)}\}.$$

From this we predict Tweety flies and is not a penguin.

Suppose we also know that

fact penguin(opus).

In this case, the communication conventions *pdf(opus)* and *bf(opus)* conflict so all theories are pruned.

Another kind of hypothesis mentioned in the previous chapter is the *conjecture* (cf. [64, 68]). Unlike defaults, which are assumed true in absence of evidence to the contrary, conjectures are simply hypotheses which may be assumed in forming an explanation or a *conditional* prediction. Any prediction that is based on a theory containing an instance of a conjecture is a conditional prediction. Let $\mathcal{C}$ be the subset of $\mathcal{H}$ containing hypotheses that are to be interpreted as conjectures. The commands to assert conjectures are:[8]

> **conjecture** *Hypothesis*. is synonymous with
>
> > **hypothesis** *Hypothesis*.
> > **meta fact** conjecture(*Hypothesis*).
>
> **conjecture** *Hypothesis*: *Wff*. is synonymous with
>
> > **hypothesis** *Hypothesis*: *Wff*.
> > **meta fact** conjecture(*Hypothesis*).

The pruning criterion for conjectures is implicit in the consistency requirement for theories given in section 3.1.2.

```
fact animal(tweety) & feathered(tweety).
fact animal(X) ← bird(X).
credulous bf(X): fly(X) ← bird(X).
credulous adf(X): ¬fly(X) ← animal(X).
conjecture fb(X): bird(X) ← feathered(X).
```

Example 3.4: "Feathered Things are Birds" Conjecture

Example 3.4 illustrates the use of conjectures. Here our knowledge is such that we are willing to entertain the possibility that feathered things are birds even though we do not necessarily accept this as plausible in the absence of evidence to the contrary. There are three maximal theories:

[8] The meta-predicate *conjecture(H)* means $H \in \mathcal{C}$.

$$T_1 = \{\mathrm{bf(tweety).adf(tweety)}\}.$$

$$T_2 = \{\mathrm{bf(tweety).fb(tweety)}\}.$$

$$T_3 = \{\mathrm{adf(tweety).fb(tweety)}\}.$$

Since these theories disagree about whether Tweety flies, it is not predicted according to our earlier definition; however, we distinguish *conditional* and *strict*[9] predictions according to whether conjectures are involved. Tweety does not fly is a strict prediction based on T_1 and if we assume *bf(tweety)* is more specific than *adf(tweety)*) so that T_3 is eliminated, then Tweety flies is a conditional prediction given *fb(tweety)* based on T_2.

Many other kinds of hypotheses can be defined in this framework. For example, we may designate some hypotheses as *askables*. Part of the pruning criterion for askables would involve a query to the user. *Fixed predicates* [65] can also be represented.

The essential point of this section is that each kind of hypothesis is simply a hypothesis plus conditions on its applicability. Meta-Theorist explicitly represents and reasons with these conditions.

[9] The term unconditional prediction is avoided since Poole [65] assigns a special meaning to this in conjunction with fixed predicates.

3.3 Meta-Theorist Interpreter

The Meta-Theorist interpreter is built upon a previous version of Theorist, Ferguson's C-Theorist [11], and is written in Quintus Prolog. The implementation is intended only as a prototype for experimentation. The purpose of this section is to describe the key aspect of the Meta-Theorist interpreter: the implementation of explanation. Other aspects of the implementation such as the user interface and the meta-level theorem prover are not discussed as they are straightforward.

The starting point for the definition of a procedure for generating explanations in Meta-Theorist is Theorem 3 (Explanation Proofs). Examination of this theorem reveals that incremental consistency checking (highlighted in Figure 5) relies on the monotonic nature of inconsistency, i.e., if a theory is inconsistent then so is any theory containing it. As Theorem 9 shows, Meta-Theorist's (non-pruned or viable) explanation procedure can be implemented by replacing the incremental consistency checking with incremental theory pruning provided the criterion is monotonic as defined below.

Definition 8 (MONOTONIC PRUNING CRITERION)

A pruning criterion $\mathcal{U}$ is *monotonic* iff for all theories T_1 and T_2 we have

$$T_1 \subseteq T_2 \;\&\; \mathcal{U}(T_1) \;\rightarrow\; \mathcal{U}(T_2).$$

Note that consistency with respect to F is a monotonic pruning criterion.

Theorem 9 (Incremental Theory Pruning)

If $\mathcal{U}$ is a monotonic pruning criterion and if $T = \{t_1, \ldots, t_n\}$ is a set of ground instances of elements of $\mathcal{H}$ then $F \cup T$ is an explanation of g iff there is a ground proof of g from $F \cup T$ such that $\neg\mathcal{U}(\{t_1, \ldots, t_i\})$ for all $i \in (1, \ldots, n)$.

Proof: If $F \cup T$ is an explanation of g then $F \cup T \models g$ and $\neg\mathcal{U}(T)$. By the completeness[10] of $\vdash$ we have $F \cup T \vdash g$. So there is a ground proof of g from $F \cup T$. Now suppose $\mathcal{U}(\{t_1, \ldots, t_i\})$ for some $i \in (1, \ldots, n)$; then, since $\{t_1, \ldots, t_i\} \subseteq T$ we have $\mathcal{U}(T)$ by the monotonicity of $\mathcal{U}$. This contradicts $\neg\mathcal{U}(T)$. Hence $\neg\mathcal{U}(\{t_1, \ldots, t_i\})$ for all $i \in (1, \ldots, n)$.

If there is a ground proof of g from $F \cup T$ then $F \cup T \models g$ by the soundness of $\vdash$. Since $\neg\mathcal{U}(\{t_1, \ldots, t_i\})$ for all $i \in (1, \ldots, n)$, in particular, we have for $i = n$ that $\neg\mathcal{U}(\{t_1, \ldots, t_n\})$, i.e., $\neg\mathcal{U}(T)$. Hence $F \cup T$ is an explanation of g. ■

Based on this theorem, we can modify the implementation of explanation given in Figure 5 by replacing the incremental consistency check highlighted in the fourth clause of **prove** with incremental theory pruning as indicated in Figure 12. This is correct in the sense that all resulting explanations are viable because pruning is checked each time an assumption is introduced (in particular, the last check is of the complete theory). It is also complete in the

[10] Our theorem prover is only complete in the sense that if $F \cup T$ is consistent and $F \cup T \models g$ then $F \cup T \vdash g$. In this case, we require the extra condition that $\neg\mathcal{U}(T)$ implies consistent($F \cup T$).

sense that no viable explanations are lost when a partially constructed theory is pruned because all theories containing a pruned subtheory are non-viable by Theorem 9.

```
prove(Goal,Ancestors,InTheory,[Goal|InTheory]) :-
    hypothesis(Goal),
    not( member(PreviousHypothesis,InTheory),
         Goal == PreviousHypothesis ),
    not( prune([Goal|InTheory]) ).
```

Figure 12: Explain with Incremental Pruning in Prolog

A similar result holds for incremental theory preference.

Definition 10 (MONOTONIC PREFERENCE CRITERION)

A preference criterion $\mathcal{P}$ (which induces the partial order $\geq^{\mathcal{P}}$ on the possible theories determined by the pruning criterion $\mathcal{U}$) is *monotonic* iff for all possible theories T_1 and T_2 we have

$$T_1 \subseteq T_2 \rightarrow T_1 \geq^{\mathcal{P}} T_2.$$

Note that subset preference (i.e., when $\geq^{\mathcal{P}}$ is $\subseteq$) underlies all monotonic preference criteria.

Theorem 11 (INCREMENTAL THEORY PREFERENCE)

If $\mathcal{P}$ is a monotonic preference criterion and if $T = \{t_1, \ldots, t_n\}$ is a set of ground instances of elements of $\mathcal{H}$ then $F \cup T$ is a preferred explanation of g

iff there is a ground proof of g from $F \cup T$ such that there is no explanation $F \cup T'$ of g where $T' >^{\mathcal{P}} \{t_1, \ldots, t_i\}$ for any $i \in (1, \ldots, n)$.

Proof: If $F \cup T$ is a preferred explanation of g then $F \cup T \models g$ and there is no explanation $F \cup T'$ of g where $T' >^{\mathcal{P}} T$. By the completeness of $\vdash$ we have $F \cup T \vdash g$. So there is a ground proof of g from $F \cup T$. Now suppose there is an explanation $F \cup T'$ of g where $T' >^{\mathcal{P}} \{t_1, \ldots, t_i\}$ for some $i \in (1, \ldots, n)$ then since $\{t_1, \ldots, t_i\} \subseteq T$ we have $T' >^{\mathcal{P}} T$ by the monotonicity of $\mathcal{P}$. This contradicts that there is no such explanation $F \cup T'$ of g where $T' >^{\mathcal{P}} T$. Hence there is no explanation $F \cup T'$ of g where $T' >^{\mathcal{P}} \{t_1, \ldots, t_i\}$ for any $i \in (1, \ldots, n)$.

If there is a ground proof of g from $F \cup T$ then $F \cup T \models g$ by the soundness of $\vdash$. Since every explanation $F \cup T'$ of g is such that $T' \not>^{\mathcal{P}} \{t_1, \ldots, t_i\}$ for all $i \in (1, \ldots, n)$, in particular, we have for $i = n$ that $T' \not>^{\mathcal{P}} T$. Hence $F \cup T$ is a preferred explanation of g. ■

As with the pruning criterion, we can implement incremental theory preference if the preference criterion is monotonic, but there is an important difference. In the case of pruning, Theorem 9 indicates that the ground proof said to exist has the property that, at every stage of partial completion, the associated partial theory is viable. The condition refers only to the partial theories involved. In the case of preference, however, the condition in Theorem 11 involves, not only the partial theories, but all other theories no other theory is strictly preferred over the partial theories. Having to check the partial theories against all other theories defeats the purpose of incremental computation – we might as well generate all explanations and then select the best theories.

Nevertheless, we can still make use of this result. In Figure 13, we again modify the highlighted part of the fourth clause of **prove**. The predicate **preferred** ensures that there is no member of a *predesignated* set of theories **CurrentBest** that is strictly preferred over the partial theory. The result is that the explanation procedure will now generate only theories that are at least as preferred as the ones designated by the predicate **currentBest**.

```
prove(Goal,Ancestors,InTheory,[Goal|InTheory]) :-
    hypothesis(Goal),
    not( member(PreviousHypothesis,InTheory),
         Goal == PreviousHypothesis ),
    preferred([Goal|InTheory]).
    viable([Goal|InTheory]).


preferred([Goal|InTheory]) :-
    currentBest(CurrentBest),
    not( member(SomeTheory,CurrentBest),
         strictly_prefer(SomeTheory,[Goal|InTheory]) ).
```

Figure 13: Explain with Incremental Preference in Prolog

We can now take this procedure and embed it in a depth-first iterative deepening search procedure. Each time the procedure produces an explanation, the **CurrentBest** list is revised (some theories may be dropped). Upon termination. `CurrentBest` contains all of the maximally preferred viable theories.

As an example of the utility of this procedure, consider the following:

$$
\begin{aligned}
F = \{\; & \mathrm{f} \;\&\; \mathrm{g} \;\&\; \mathrm{h} \rightarrow \mathrm{k}, \\
& \mathrm{a} \;\&\; \mathrm{b} \rightarrow \mathrm{k}, \\
& \mathrm{a} \rightarrow \mathrm{f}, \\
& \mathrm{b} \;\&\; \mathrm{c} \rightarrow \mathrm{g}, \\
& \mathrm{l}_n,\ \mathrm{l}_n \rightarrow \mathrm{l}_{n-1},\ \ldots,\ \mathrm{l}_2 \rightarrow \mathrm{l}_1,\ \mathrm{l}_1 \rightarrow \mathrm{h} \;\} \\
\mathcal{H} = \{\; & \mathrm{a},\ \mathrm{b},\ \mathrm{c} \;\}
\end{aligned}
$$

Suppose we take the pruning criterion to be consistency and the preference criterion to be subset preference. There are two viable theories explaining k: $T_1 = \{$ a, b, c $\}$ and $T_2 = \{$ a, b $\}$. Theory T_2 is strictly preferred over theory T_1 because it is a proper subset. Note that computing all viable theories involves the arbitrarily expensive proof of h. This means that finding all preferred viable theories by first finding all viable theories is arbitrarily expensive.

Our procedure, however, finds the best theory in two iterations while visiting only 12 nodes (4 at depth 1 and 8 at depth 2). After the first iteration, a proof of k from the theory T_2 is found and it becomes the current best theory. On the second iteration, we encounter c in the sub-proof of g having previously assumed a and b. At this point, the sub-proof of g is abandoned because the current best theory is strictly preferred to the partial theory { a, b, c }. Since there were no branches cutoff by depth considerations during this iteration, the procedure halts with the current best theory T_2.

The incremental pruning and preference theorems only guarantee the correctness of the given implementations when the pruning and preference criteria are monotonic. Unfortunately, not all of the pruning and preference criteria we are interested in are monotonic. For instance, in section 3.2 it was mentioned that we may want our communication conventions to be relative to the facts plus any conjectures we make. Pruning criteria for such conventions will be nonmonotonic because conjectures might enable conflicts between conventions. It can happen that a convention that was unblocked when we assumed it becomes blocked when later assuming a conjecture.

There are at least three strategies that can be used to deal with nonmonotonic pruning and preference criteria. First, we could delay checking some hypotheses. For instance, we could incrementally check conjectures and delay checking conventions until we reach a point where no further conjectures will be assumed. Some versions of Prolog use such a delaying strategy to deal with unbound variables in negation [54]. Poole uses a similar technique to deal with unbound variables in consistency checking [67].

Second, we could recheck some hypotheses. For instance, every time we add a conjecture, we could recheck all the conventions in the current theory to see if any previously unblocked conventions have become blocked.

Third, we can vary the size of the checking increment. For instance, when considering adding a convention, if we find it is currently blocked, we could consider adding in conjectures to unblock the convention. The effect of this is that a pruning criterion that is nonmonotonic for single steps becomes monotonic between variable sized steps. We call such pruning and preference criteria *stratified monotonic* criteria. This reflects that the pruning criteria can be divided into multiple layers. The bottom layer consists of a monotonic criterion, e.g., we might consider only whether hypotheses are consistent conjectures. The criterion at the next layer becomes monotonic once the bottom layer is fixed, e.g., if no more conjectures are considered, the criterion for conventions becomes monotonic.

So far we have only considered explanation implementation issues. When prediction was described earlier, it was in terms of the consequences of all maximal theories. This definition of prediction presupposes a particular interpretation of defaults, namely, credulous defaults. In the remainder of this dissertation, a different interpretation of defaults is developed. As will be seen, under this interpretation, prediction becomes simply conjecture-free explanation.

3.4 Conclusion

This chapter has presented an extension to Theorist called Meta-Theorist. Meta-Theorist augments Theorist with a meta-level in which theory pruning and theory preference may be specified. Consequently, Meta-Theorist is capable of a broader range of representation and reasoning than Theorist. For instance, Meta-Theorist can reason about preferred explanations while Theorist is limited to consistent explanations. Meta-Theorist is a general tool for common sense reasoning. The meta-level allows specification of theory pruning and theory preference in first order logic. This has the advantage of expressiveness but the disadvantage of providing the user with little guidance and being computationally expensive. However, Meta-Theorist was developed simply to allow experimentation with various pruning and preference criteria [illegible] adequate for this task.

Chapter 4

Probabilistic Knowledge and Inference

It has been argued that the nature of much of our knowledge of the world and much of our reasoning is probabilistic, and therefore, probabilistic knowledge and inference is a crucial component of the knowledge representation and reasoning problem in AI (e.g., [51]). This chapter discusses probabilistic knowledge and inference; in particular, the combined probability logic developed by Bacchus[3] and Halpern [19] is examined.[1] A fundamental problem in probabilistic inference—the reference class problem—is also described. This chapter provides the necessary background for the next chapter where we consider the reference class problem more closely.

[1] Except where noted, mention of Bacchus's work in this chapter refers to [3].

4.1 Probabilities

Like logic, probabilities have been intensely studied. Simple axiomatizations, e.g., [31], are well understood and generally accepted. Yet much debate remains over the interpretation (and application) of probabilities. The axiomatization of probabilities restricts the possible interpretations to satisfy certain properties, but the axiomatization does not pick out a unique interpretation, nor does the axiomatization interpret itself.

Three main interpretations of probabilities have been proposed. the empirical, the logical, and the subjective [illegible] The *empirical* interpretation views probability statements as statistical claims about the domain These claims have an objective truth value that depends only on the state of the domain and not on any body of evidence or on any agent's beliefs. The *logical* interpretation takes probabilities to be a predetermined logical relation between an assertion and a body of evidence. In the *subjective* interpretation, probabilities are understood as the degrees of belief of some agent at some time. These degrees of belief may represent opinions that are not necessarily grounded in reality.

Carnap [5] has suggested the need for two distinct kinds of probabilities (this view was recently adopted for AI by Bacchus and Halpern [3, 19].). One kind of probability is needed for empirical knowledge and another is needed for beliefs. Consider the following two statements:

Birds probably fly.

Tweety (a particular bird) probably flies.

Bacchus has shown that straightforward approaches to representing the above two statements in a single probability interpretation run into difficulties.

For example, if we interpret probabilities empirically, the first statement is true in the domain if most of the individuals in the domain that are birds also fly. But the second statement presents a problem since Tweety is a particular individual in the domain who either flies or doesn't fly. That is, the probability that Tweety flies under the empirical interpretation is either zero or one. The empirical interpretation does not capture the distinction between "Tweety probably flies" and "Tweety flies."

If instead we interpret probabilities subjectively, say with a possible world semantics such as [56], Bacchus has shown that straightforward approaches to representing the first statement do not work. For instance, attaching a high probability to the statement

$$\forall \texttt{X.bird(X)} \rightarrow \texttt{fly(X)}$$

is inadequate since this means that all birds fly in most possible worlds rather than most birds fly in all possible worlds. This distinction is important since the latter allows for the possibility of there being at least one non-flying bird in every possible world while the former rules this out.

The work Bacchus [3] and Halpern [19] has yielded a combined probability logic capable of representing and reasoning with these two kinds of probabilities. In the next two sections, we discuss (empirical) probabilities on the domain and (subjective) probabilities on possible worlds. Bacchus refers to these two types of probabilities as *statistical* and *propositional* probabilities respectively.

4.2 Statistical Probabilities

Much of our knowledge is essentially statistical knowledge. Statistical probabilities are probabilities defined over sets of individuals. They make a claim about the domain which is either true or false. Statistical knowledge is empirically verifiable, though in practice, verifiability is limited by our observational powers.

Examples of statistical probabilities are: the probability that birds fly, the probability that elephants like zookeepers, the probability that adults are employed.

Bacchus develops a syntax, semantics, and proof theory for a two-sorted first order logic containing statistical terms. The approach involves placing a probability distribution on the domain of discourse $\mathcal{O}$ to assign probabilities to sets of individuals. Formulas with placeholder variables are used to denote sets of individuals or vectors of individuals. Numeric terms denoting the probability of the set of individuals defined by a formula are generated using a variable binding statistical probability operator. This operator consists of square brackets, subscripted by a set of placeholder variables, placed around its formula argument. For example, $[\texttt{bird(X)}]_{\texttt{X}}$ represents[2] the measure of the set of birds. Conditional probabilities are introduced as a definitional extension to the logic. The conditional probability, for instance, of fly given bird is represented as $[\texttt{fly(X)}|\texttt{bird(X)}]_{\texttt{X}}$. More generally, the notation

$$[\texttt{p}(\vec{X})|\texttt{q}(\vec{X})]_{\vec{X}}$$

is used to express the measure of the set of vectors of domain objects $\vec{X}$ that satisfies $\texttt{p}(\vec{X})$ given they satisfy $\texttt{q}(\vec{X})$.

Below are the statistical probability terms corresponding to the three examples given above.

1. $[\texttt{fly(X)}|\texttt{bird(X)}]_{\texttt{X}}$.

2. $[\texttt{likes(X,Y)}|\texttt{elephant(X)} \ \& \ \texttt{zookeeper(Y)}]_{(\texttt{X},\texttt{Y})}$.

3. $[\texttt{employed(X)}|\texttt{adult(X)}]_{\texttt{X}}$.

Assertions about these statistical probabilities can be made in the logic. For instance, "More than 75% of all birds fly," "90% of pairs of elephants and zookeepers are in the likes relation," "More adults are employed than unemployed" correspond respectively to

[2] For convenience, we use the Prolog convention of upper case variables and lower case constants and predicate names. The reverse convention is used by Bacchus.

1. $[\texttt{fly(X)}|\texttt{bird(X)}]_{\texttt{X}} > 0.75$.

2. $[\texttt{likes(X,Y)}|\texttt{elephant(X) \& zookeeper(Y)}]_{(\texttt{X},\texttt{Y})} = 0.9$.

3. $[\texttt{employed(X)}|\texttt{adult(X)}]_{\texttt{X}} > [\neg\texttt{employed(X)}|\texttt{adult(X)}]_{\texttt{X}}$.

   or, equivalently,

   $[\texttt{employed(X)}|\texttt{adult(X)}]_{\texttt{X}} > 0.5$.

A discussion of the formal details of Bacchus's logic is unnecessary for our purposes; however, we give several useful lemmas. These lemmas show that certain probability terms can be simplified or related to simpler probability terms.

Lemma 12 (Bacchus)

If no $x_i \in \vec{x}$ which appears in $\alpha \wedge \beta$ is free in λ then

$$\models [\beta \wedge \lambda]_{\vec{x}} > 0 \rightarrow [\alpha|\beta \wedge \lambda]_{\vec{x}} = [\alpha|\beta]_{\vec{x}}.$$

Two special cases of the above lemma are given in the following corollary.

Corollary 13 (Bacchus)

1. If no $x_i \in \vec{x}$ is free in λ then

$$\models [\beta \wedge \lambda]_{\vec{x}} > 0 \rightarrow [\alpha|\beta \wedge \lambda]_{\vec{x}} = [\alpha|\beta]_{\vec{x}}.$$

2. If α and λ share no free variables then

$$\models [\lambda]_{\vec{x}} > 0 \rightarrow [\alpha|\lambda]_{\vec{x}} = [\alpha]_{\vec{x}}.$$

For instance, if $[\texttt{fly(X)|bird(X) \& dog(Y)}]_{\langle \texttt{X,Y} \rangle} > 0$ then

$$[\texttt{fly(X)|bird(X) \& dog(Y)}]_{\langle \texttt{X,Y} \rangle} = [\texttt{fly(X)|bird(X)}]_{\langle \texttt{X,Y} \rangle}.$$

Lemma 14 (Bacchus)

If y does not appear free in α then

$$\models [\alpha]_{<\vec{x},y>} = [\alpha]_{\vec{x}}.$$

So, for example, by Lemma 14 and the definition of conditional probability terms, we have $[\texttt{fly(X)|bird(X)}]_{\langle \texttt{X,Y} \rangle} = [\texttt{fly(X)|bird(X)}]_{\texttt{X}}$.

Lemma 15 (Bacchus)

$$\models \forall \vec{x}.(\beta \rightarrow \lambda) \rightarrow [\alpha|\beta \wedge \lambda]_{\vec{x}} = [\alpha|\beta]_{\vec{x}}.$$

Lemma 16 (Bacchus)

$$\models \forall \vec{x}.(\alpha \rightarrow \lambda) \rightarrow [\lambda|\beta]_{\vec{x}} \geq [\alpha|\beta]_{\vec{x}}.$$

The following two lemmas (and, more specifically, Lemma 39 in Chapter 6) are useful for the kind of reasoning involved in cases such as the Elephants and Zookeepers problem (Example 6.2 in Chapter 6).

Lemma 17 (Bacchus)

$$\models \quad \forall r_1 r_2. \Big[[\alpha|\beta]_{\vec{x}} \in [r_1, r_2] \wedge \beta\Big]_{\langle \vec{x},\vec{y} \rangle} > 0$$
$$\rightarrow \Big[\alpha\Big|[\alpha|\beta]_{\vec{x}} \in [r_1, r_2] \wedge \beta\Big]_{\langle \vec{x},\vec{y} \rangle} \in [r_1, r_2].$$

Lemma 18 (BACCHUS)

If no $x_i \in \vec{x}$ is free in the formula δ then

$$\models \forall r_1 r_2 . \Big[\beta \wedge \delta \wedge [\alpha|\beta]_{\vec{x}} \in [r_1, r_2]\Big]_{\langle \vec{x}, \vec{y} \rangle} > 0$$
$$\rightarrow \Big[\alpha \Big| \beta \wedge \delta \wedge [\alpha|\beta]_{\vec{x}} \in [r_1, r_2]\Big]_{\langle \vec{x}, \vec{y} \rangle} \in [r_1, r_2].$$

Lemma 19 (BACCHUS)

If α is a closed formula, i.e., if α has no free variables, then $[\alpha]_{\vec{x}} = 0$ or 1.

This final lemma shows that statistical probabilities are unsuitable for representing propositional probabilities. For instance, $[fly(tweety)]_{\vec{x}} = 0$ or 1.

4.3 Propositional Probabilities

Propositional probabilities are assertions about the subjective state of an agent's beliefs. They express an agent's degree of belief in some proposition. The agent to which reference is being made is left implicit so, for instance, the degree of belief in some proposition is understood to mean that particular agent's degree of belief in that proposition. Both the assertion that a particular agent holds a particular degree of belief in some proposition and the proposition itself are objectively true or false, but the agent's degree of belief in some proposition is not. For example, even though the world is in fact not flat, young Randy, from Melville Saskatchewan may believe that it is flat. So his degree of belief in the proposition "the earth is flat" is 1. Propositional probabilities may be unconnected to the domain, yet such a connection is desirable as is discussed in the next section.

Examples of propositional probabilities are: the probability of the proposition "Tweety flies" given "Tweety is a bird," the probability of the proposition "Clyde likes Tony" given "Clyde is an elephant and Tony is a zookeeper," the probability of the proposition "Randy is employed" given "Randy is an adult."

Halpern [19] has developed a probability logic that combines Bacchus's statistical logic with propositional probabilities. This logic includes a propositional probability operator that maps formulas to the probability of the set of possible worlds satisfying the formula. The syntax $\mathsf{prob}(\alpha)$ denotes the propositional probability of the formula α, i.e., the probability of the set of worlds satisfying α. Conditional probabilities are introduced by definitional extension of the logic. The notation

$$\mathsf{prob}\big(\mathsf{p}(\vec{a})|\mathsf{q}(\vec{a})\big)$$

is used to express the propositional probability that a particular vector of domain objects $\vec{a}$ satisfies p given that it satisfies q. Another definitional extension is an abbreviation for probability one.

$$\mathsf{cert}(\alpha) =_{df} \mathsf{prob}(\alpha) = 1.$$

Propositional probability terms corresponding to the three examples given above are

1. $\mathrm{prob}\big(\texttt{fly(tweety)}|\texttt{bird(tweety)}\big)$.

2. $\mathrm{prob}\big(\texttt{likes(clyde,tony)}|\texttt{elephant(clyde)}\ \&\ \texttt{zookeeper(tony)}\big)$.

3. $\mathrm{prob}\big(\texttt{employed(randy)}|\texttt{adult(randy)}\big)$.

Assertions about these propositional probabilities can be made in the logic. For instance, the probability of the proposition "Tweety flies" given "Tweety is a bird" is more than 75%, the probability of the proposition "Clyde likes Tony" given "Clyde is an elephant and Tony is a zookeeper" is 90%, the probability of the proposition "Randy is employed" given "Randy is an adult" is greater than the probability of the proposition "Randy is unemployed" given "Randy is an adult" correspond respectively to

1. $\mathrm{prob}\big(\texttt{fly(tweety)}|\texttt{bird(tweety)}\big) > 0.75$,

2. $\mathrm{prob}\big(\texttt{likes(clyde,tony)}|\texttt{elephant(clyde)}\ \&\ \texttt{zookeeper(tony)}\big)$
   $= 0.9$,

3. $\mathrm{prob}\big(\texttt{employed(randy)}|\texttt{adult(randy)}\big)$
   $> \mathrm{prob}\big(\neg\texttt{employed(randy)}|\texttt{adult(randy)}\big)$,
   or, equivalently,
   $\mathrm{prob}\big(\texttt{employed(randy)}|\texttt{adult(randy)}\big) > 0.5$.

Bacchus argues that propositional probabilities are inadequate for representing statistical probabilities. Though there are technical ways of representing these, the obvious schemes do not work. For example, representing "More than 75% of all birds fly" as a high probability universal, i.e.

$$\mathsf{prob}\Big(\forall \mathtt{X}.\mathtt{bird(X)} \rightarrow \mathtt{fly(X)}\Big) > 0.75$$

precludes assigning a high probability to there being a non-flying bird since in every possible world where there is a non-flying bird, the universal is false. This is clearly seen in the equality

$$\mathsf{prob}\Big(\forall \mathtt{X}.\mathtt{bird(X)} \rightarrow \mathtt{fly(X)}\Big) = 1 - \mathsf{prob}\Big(\exists \mathtt{X}.\mathtt{bird(X)}\ \&\ \neg\mathtt{fly(X)}\Big).$$

Another possibility is to represent the statistical probability as

$$\forall \mathtt{X}.\mathsf{prob}\Big(\mathtt{bird(X)}\Big) > 0 \rightarrow \mathsf{prob}\Big(\mathtt{fly(X)}|\mathtt{bird(X)}\Big) > 0.75.$$

This is also inadequate because it precludes believing there is that a particular bird than does not fly. For instance, the above assertion is inconsistent with

$$\mathsf{prob}\Big(\mathtt{bird(opus)}\Big) > 0 \rightarrow \mathsf{cert}\Big(\neg\mathtt{fly(opus)}|\mathtt{bird(opus)}\Big).$$

Because both statistical and propositional probabilities are needed for representing and reasoning about our probabilistic knowledge, Bacchus combines the two kinds of probabilities in one logic. This logic is capable of expressing assertions with both kinds of probabilities and reasoning about their valid consequences. But because the probability distribution over the domain does not restrict the probability distribution over the possible worlds, the two kinds of probabilities are unconnected. The next section examines the problem of establishing a connection between statistical and propositional probabilities.

4.4 Direct Inference

> "The issue of empirical foundations for probabilities *used as degrees of belief* has been largely ignored in AI. Proponents of probabilities have always been quick to claim that probabilities are empirically founded, referring to statistical probabilities. Unfortunately they often continue on to claim that propositional probabilities have the same advantage, without paying due attention to the serious problems involved in connecting propositional probabilities with empirical observations." [3, p. 141]

If propositional probabilities are not empirically founded, where do they come from and how can we judge whether they are rational? If they are empirically founded, then how?

As Bacchus has noted, a connection is made between propositional probabilities and statistical probabilities in actuarial situations. The rate quoted for a particular person's life insurance is based on assuming the propositional probability of the person's dying (say, this year) given various things that are known about the person, such as age and occupation, is equal to the proportion of deaths (statistical probability) among the group of individuals having these properties.

In philosophy, the inference of propositional probabilities from statistical probabilities is an idea going back at least to Reichenbach [74]. He suggested that single case probabilities (propositional) should be determined by consulting the statistical probabilities for the narrowest class to which the single case belongs and for which there are "adequate statistics."

For instance, we infer that the propositional probability that the bird Tweety flies is equal to the statistical probability that birds fly provided we know this statistic and we do not know the statistical probability of flying for any smaller class to which Tweety belongs. If Tweety was also known to be yellow then yellow birds is a smaller class to which Tweety is known to belong. Reichenbach's principle still allows us to infer the propositional probability of Tweety flying from the statistical probability of birds flying since we have no statistics for yellow birds. This amounts to assuming that yellowness is irrelevant to Tweety's flying abilities.

Inferring propositional probabilities from statistical probabilities is called *direct inference* [61]. This inference depends on choosing a *reference class* which contains the single case. The *reference class problem* is the problem of choosing the appropriate class from which to perform direct inference [35]. A number of techniques or policies for making this choice have been proposed, e.g., [34, 39, 35, 60, 61, 3].

Bacchus defines a *direct inference principle* in his combined probability logic to connect statistical and propositional probabilities. He starts by assuming that an agent expresses assertions about his environment in a fixed statistical language $\mathcal{L}^{\text{stat}}$. Assertions in $\mathcal{L}^{\text{stat}}$ are called *objective* assertions. The agent's degree of belief in the objective assertions are represented in another language $\mathcal{L}^{\text{comb}}$ which extends $\mathcal{L}^{\text{stat}}$ with the propositional probability operator prob and an expectation operator E. (The expectation operator maps terms to their expected value, i.e., the weighed average of their value across possible worlds.) Formulas of $\mathcal{L}^{\text{comb}}$ that are also in $\mathcal{L}^{\text{stat}}$ are called *objective* formulas. The agent's knowledge base KB is a finite collection of objective formulas. KB represents the agent's full beliefs, i.e., cert(KB). The propositional probabilities are determined by direct inference from KB. Example 4.1 illustrates direct inference in Bacchus's logic.

The following lemma summarizes many useful properties of the expectation operator.

Lemma 20 (PROPERTIES OF THE EXPECTATION TERMS (BACCHUS))

1. If t is rigid,[3] then $\mathsf{E}(t) = t$.

[3] Rigid terms are terms whose denotation is invariant across possible worlds, e.g., numeric constants.

2. $\mathsf{cert}(t = t') \rightarrow \mathsf{E}(t) = \mathsf{E}(t')$.
3. $\mathsf{cert}(t < t') \rightarrow \mathsf{E}(t) < \mathsf{E}(t')$.
4. $\mathsf{cert}(t > t') \rightarrow \mathsf{E}(t) > \mathsf{E}(t')$.
5. $\mathsf{cert}(t \leq t') \rightarrow \mathsf{E}(t) \leq \mathsf{E}(t')$.
6. $\mathsf{cert}(t \geq t') \rightarrow \mathsf{E}(t) \geq \mathsf{E}(t')$.
7. $\mathsf{E}(t) = 0 \equiv \mathsf{cert}(t = 0)$.
8. $\mathsf{E}(t) = 1 \equiv \mathsf{cert}(t = 1)$.
9. $\mathsf{E}(t + t') = \mathsf{E}(t) + \mathsf{E}(t')$.
10. If r is rigid, then $\mathsf{E}(r \times t) = r \times \mathsf{E}(t)$.
11. If r is rigid, then $\mathsf{E}(t/r) = \mathsf{E}(t)/r$.

Prior to formally defining direct inference, the concept of randomization must be defined.

Definition 21 (RANDOMIZATION (BACCHUS))

Let α be a formula of $\mathcal{L}^{\mathsf{stat}}$. If $\langle c_1, \ldots, c_n \rangle$ are the n distinct object constants that appear in $\alpha \wedge \mathsf{KB}$ and $\langle v_1, \ldots, v_n \rangle$ are n distinct object variables that do not occur in $\alpha \wedge \mathsf{KB}$, then let KB^{v} (α^{v}) denote the new formula which results from textually substituting c_i by v_i in KB (α), for all i.

Now we can give the definition of direct inference which links the two kinds of probabilities.

Definition 22 (DIRECT INFERENCE PRINCIPLE (BACCHUS))

If α is a formula of $\mathcal{L}^{\mathsf{stat}}$ and if KB is the complete set of objective formulas that the agent fully believes, then the agent's degree of belief in α should be determined by the equality

$$\mathsf{prob}(\alpha) = \mathsf{E}\Big([\alpha^{\mathsf{v}}|\mathsf{KB}^{\mathsf{v}}]_{\vec{v}}\Big).$$

The agent must also fully believe that $[\mathsf{KB}^{\mathsf{v}}]_{\vec{v}} > 0$, i.e., $\mathsf{cert}\Big([\mathsf{KB}^{\mathsf{v}}]_{\vec{v}} > 0\Big)$.

The above definition leads to the property that an agent fully believes all the consequences of his full beliefs. This *logical omniscience* property is controversial (e.g., [10]).

Theorem 23 (BACCHUS)

$\mathsf{prob}(\alpha) = 1$ is a logical consequence of the direct inference principle if and only if $\mathsf{KB} \models \alpha$.

The theory generated by the direct inference principle is formally defined as follows:

Definition 24 (THE THEORY T_0 (BACCHUS))

Let D_0 be a set of formulas of $\mathcal{L}^{\mathsf{comb}}$ consisting of the formula $\mathsf{cert}\Big([\mathsf{KB}^{\mathsf{v}}]_{\vec{v}}$ $> 0\Big)$ along with all instances of the direct inference principle. That is, for all objective formulas α, i.e., $\alpha \in \mathcal{L}^{\mathsf{stat}}$, D_0 will contain the formula $\mathsf{prob}(\alpha) =$ $\mathsf{E}\Big([\alpha^{\mathsf{v}}|\mathsf{KB}^{\mathsf{v}}]_{\vec{v}}\Big)$. Let T_0 denote the *closure of D_0 under logical consequence.*

Let

$$\mathsf{KB} = \mathtt{bird(tweety)}\ \&\ [\mathtt{fly(X)}|\mathtt{bird(X)}]_{\mathtt{X}} > c,$$

where c is a *rigid* numeric constant that is strictly greater than 0.5. Then we have the following proof:

$$\begin{aligned}
&\mathsf{prob}\Big(\mathtt{fly(tweety)}\Big)\\
&\qquad = \mathsf{E}\Big(\Big[\mathtt{fly(V)}\Big|\mathtt{bird(V)}\ \&\ [\mathtt{fly(X)}|\mathtt{bird(X)}]_{\mathtt{X}} > c\Big]_{\mathtt{V}}\Big) && \text{Def. 22}\\
&\mathsf{cert}\Big(\Big[\mathtt{fly(V)}\Big|\mathtt{bird(V)}\ \&\ [\mathtt{fly(X)}|\mathtt{bird(X)}]_{\mathtt{X}} > c\Big]_{\mathtt{V}}\\
&\qquad = [\mathtt{fly(V)}|\mathtt{bird(V)}]_{\mathtt{V}}\Big) && \text{Cor. 13}\\
&\mathsf{E}\Big(\Big[\mathtt{fly(V)}\Big|\mathtt{bird(V)}\ \&\ [\mathtt{fly(X)}|\mathtt{bird(X)}]_{\mathtt{X}} > c\Big]_{\mathtt{V}}\Big)\\
&\qquad = \mathsf{E}\Big([\mathtt{fly(V)}|\mathtt{bird(V)}]_{\mathtt{V}}\Big) && \text{Lem. 20.2}\\
&\mathsf{prob}\Big(\mathtt{fly(tweety)}\Big) = \mathsf{E}\Big([\mathtt{fly(V)}|\mathtt{bird(V)}]_{\mathtt{V}}\Big)\\
&\mathsf{cert}\Big([\mathtt{fly(V)}|\mathtt{bird(V)}]_{\mathtt{V}} > c\Big) && \text{Thm. 23}\\
&\mathsf{E}\Big([\mathtt{fly(V)}|\mathtt{bird(V)}]_{\mathtt{V}}\Big) > \mathsf{E}(c) && \text{Lem. 20.4}\\
&\mathsf{E}(c) = c && \text{Lem. 20.1}\\
&\mathsf{prob}\Big(\mathtt{fly(tweety)}\Big) > c
\end{aligned}$$

Example 4.1: Tweety flies (Bacchus)

4.5 Inadequate Statistics (Reference Class Problem)

The direct inference principle in Bacchus's combined probability logic takes into account all that is known about the individuals in the formula to which it is applied. It uses the narrowest reference class—a reference class which may be so narrow that we have no useful statistics about it.

SKB4.2:

$$\begin{aligned}&[\texttt{fly(X)}|\texttt{bird(X)}]_{\texttt{X}} = 0.75,\\ &[\texttt{fly(X)}|\texttt{antarctic(X)}\ \&\ \texttt{bird(X)}]_{\texttt{X}} = 0.2,\\ &\texttt{black(opus)}\ \&\ \texttt{antarctic(opus)}\ \&\ \texttt{bird(opus)}\end{aligned}$$

Example 4.2: Black Antarctic Birds

Suppose the agent's statistical KB is the the conjunction of the formulas listed in SKB4.2 (Example 4.2). Using Bacchus's direct inference principle we have T_0 contains

$$\begin{aligned}&\mathsf{prob}\Big(\texttt{fly(opus)}\Big)\\ &= \mathsf{E}\Big([\texttt{fly(V)}|\texttt{black(V)}\ \&\ \texttt{antarctic(V)}\ \&\ \texttt{bird(V)}]_{\texttt{V}}\Big).\end{aligned}$$

Since no useful information about the expected value for black antarctic birds can be derived from SKB4.2, T_0 only entails

$$0 \leq \mathsf{prob}\Big(\texttt{fly(opus)}\Big) \leq 1.$$

The direct inference principle tells us nothing useful in this case.

Since the statistics for black antarctic birds are "inadequate," we should, according to Reichenbach's principle, turn to the wider reference class of antarctic birds for which SKB4.2 does entail useful information about the expected value. This amounts to assuming the irrelevance of blackness to the flying ability of antarctic birds.

Bacchus provides a mechanism for moving to the wider reference class. The next section discusses this mechanism.

4.6 Expectation Independence Assumptions

Conditioning on the entire knowledge base amounts to using the narrowest possible reference class. This reference class may be so narrow that we have no useful statistical information about it. To deal with this problem of *overly specific reference classes*, Bacchus provides a mechanism which allows the nonmonotonic inheritance of statistical information from superclasses. These assumptions are added to T_0 (Definition 24) which is the base theory resulting from direct inference.

The general form of these assumptions is

$$\mathsf{cert}(\forall\vec{v}.\mathsf{KB}^{\mathsf{v}} \rightarrow \beta) \rightarrow \mathsf{E}\Big([\alpha|\mathsf{KB}^{\mathsf{v}}]_{\vec{v}}\Big) = \mathsf{E}\Big([\alpha|\beta]_{\vec{v}}\Big).$$

These *expectation independence assumptions* are weaker than the corresponding statistical independence assumptions:

$$\mathsf{cert}(\forall\vec{v}.\mathsf{KB}^{\mathsf{v}} \rightarrow \beta) \rightarrow \mathsf{cert}\Big([\alpha|\mathsf{KB}^{\mathsf{v}}]_{\vec{v}} = [\alpha|\beta]_{\vec{v}}\Big).$$

Bacchus justifies the use of the weaker expectation independence assumptions by illustrating the *unit reference class problem* (see [3, p. 152]).

The mechanism allows, for instance in Example 4.2, moving to the wider reference class of antarctic birds by the nonmonotonic assumption

$$\begin{aligned}&\mathsf{E}\Big([\texttt{fly(V)}|\texttt{black(V)}\ \&\ \texttt{antarctic(V)}\ \&\ \texttt{bird(V)}]_{\mathtt{V}}\Big)\\&=\mathsf{E}\Big([\texttt{fly(V)}|\texttt{antarctic(V)}\ \&\ \texttt{bird(V)}]_{\mathtt{V}}\Big)\end{aligned}$$

from which we could derive

$$\mathsf{prob}\Big(\texttt{fly(opus)}\Big)=0.2.$$

The same mechanism allows moving to the wider reference class of birds by the assumption

$$\begin{aligned}&\mathsf{E}\Big([\texttt{fly(V)}|\texttt{black(V)}\ \&\ \texttt{antarctic(V)}\ \&\ \texttt{bird(V)}]_{\mathtt{V}}\Big)\\&=\mathsf{E}\Big([\texttt{fly(V)}|\texttt{bird(V)}]_{\mathtt{V}}\Big)\end{aligned}$$

from which we could derive

$$\mathsf{prob}\Big(\texttt{fly(opus)}\Big)=0.75.$$

The expectation independence assumptions permitted by Bacchus are determined by the following:

Definition 25 (NONMONOTONIC ASSUMPTIONS (BACCHUS))

If $\mathsf{cert}(\forall\vec{v}.\mathsf{KB}^{\mathsf{v}}\rightarrow\beta)\in T_0$, then

$$\mathsf{E}\Big([\alpha|\mathsf{KB}^{\mathsf{v}}]_{\vec{v}}\Big)=\mathsf{E}\Big([\alpha|\beta]_{\vec{v}}\Big)$$

is a legitimate nonmonotonic assumption.

The above definition gives rise to multiple possible extensions to the base theory T_0. These are formally described by the following definition.

Definition 26 (NONMONOTONIC THEORIES (BACCHUS))
Let T_i be the closure under logical consequence of $T_0 \cup D_i$, where D_i is the i-th finite set of default assumptions. That is, $T_i = \{\alpha : T_0 \cup D_i \models \alpha\}$.

Some of the nonmonotonic theories may be inconsistent as indicated in the following definition.

Definition 27 (CONTRADICTED THEORIES (BACCHUS))
A theory T_i *contradicts* a theory T_j if there exists a formula α such that $\alpha \in T_i$ and $\neg\alpha \in T_j$.

The following two definition characterize preferred assumptions and preferred theories.

Definition 28 (PREFERRED NONMONOTONIC ASSUMPTIONS (BACCHUS))

The nonmonotonic assumption

$$\mathsf{E}\Big([\alpha|\mathsf{KB}^{\mathsf{v}}]_{\vec{v}}\Big) = \mathsf{E}\Big([\alpha|\beta_1]_{\vec{v}}\Big)$$

is *preferred* to the nonmonotonic assumption

$$\mathsf{E}\Big([\alpha|\mathsf{KB}^{\mathsf{v}}]_{\vec{v}}\Big) = \mathsf{E}\Big([\alpha|\beta_2]_{\vec{v}}\Big).$$

if $\mathbf{cert}(\forall\vec{v}.\beta_1 \rightarrow \beta_2) \in T_0$. That is, we prefer to inherit expectations from *known* narrower classes.

Definition 29 (PREFERRED THEORIES (BACCHUS))

The theory T_i is *preferred* to the theory T_j if for every default assumption $d \in D_i$, there exists a default assumption $d' \in D_j$ such that d is preferred to d'.

Only some of the theories are considered viable descriptions of an agent's beliefs. These theories are determined by the following definition.

Definition 30 (VIABLE THEORIES (BACCHUS))

T_i is a *viable* if it is not contradicted by any preferred theory.

Bacchus's preference ordering on theories captures the preference for inheriting statistical information from narrower reference classes. Returning to Example 4.2, because antarctic birds are a subset of birds, the theory T_1 based on the first assumption is preferred to the theory T_2 based on the second assumption (by Definition 29). Furthermore, by Definition 27, T_2 is contradicted by T_1 since the two theories disagree on the value of $\mathbf{prob}\big(\texttt{fly(opus)}\big)$. Therefore T_2 is not a *viable* theory by Definition 30. Both T_0 and T_1 are viable theories for describing the agent's beliefs. Bacchus's formalism does not uniquely determine the agent's theory.

Bacchus argues (as does Konolige and Myers [32]) that domain independent determination of a unique theory is too much to expect. However, Bacchus claims:

> "What a default reasoning system should provide is the ability [to] generate the obvious, uncontroversial theories, and equally important, it should eliminate obviously incorrect theories." [3, p. 168]

In the next chapter, it is argued that Bacchus's formalism does not go far enough in eliminating obviously incorrect theories and a novel reference class selection policy based on averaging is proposed.

Chapter 5

Reference Class Selection

> "If we are asked to find the probability holding for an individual future event, we must first incorporate the case in a suitable reference class. An individual thing or event may be incorporated in many reference classes from which different probabilities will result. This ambiguity has been called the *problem of the reference class.*" [74, p. 375]

This chapter examines a fundamental problem in probabilistic inference—the reference class problem [35]—within the context of the combined probability logic developed by Bacchus [2, 3] and Halpern [19]. We begin by considering the reference class selection policy in Bacchus's formalism. This leads to the claim that Bacchus's viable theories are too liberal in some cases and too conservative in others. The difficulty lies in the "expectation independence" assumptions allowed in his formalism—the culprit being the restriction to ***downward inheritance*** of statistics.

To illustrate, suppose we know most vaccinated people acquire immunity, but blonde vaccinated children do not acquire immunity. As we argue in section 5.1, this situation is ambiguous w.r.t. vaccinated children acquiring immunity because the class vaccinated children is wedged between the conflicting classes vaccinated people and blonde vaccinated children. But Bacchus's inheritance mechanism does not consider the upwards influence of statistics for blonde vaccinated children with the consequence that most vaccinated children are presumed to acquire immunity.

We propose an alternative reference class selection policy in which the probability of a proposition concerning a particular individual is inferred from the corresponding statistic within the unique reference class of individuals that are KB-indistinguishable[1] from the particular individual. This statistic is in turn inferred from the average value of the statistic over all sets of individuals that are KB-indistinguishable from the reference class. We will show how this randomizing[2] of individuals and sets (relations) solves problems involving conflicting sources of statistical knowledge that present difficulties for many other approaches.

Two advantages of this policy are that a unique reference class is always determined and that confining the "magic" to randomization reduces the

[1] If in every interpretation that satisfies KB, exchanging two individuals i_1 and i_2 produces another interpretation that satisfies KB, then i_1 and i_2 are KB-indistinguishable.

[2] Kyburg has pointed out the potential circularity suggested in the term randomization. Perhaps meanification or averagifying or even variablizing would be more appropriate terms.

determination of relevant statistics to a straightforward application of statistical principles. Randomization involves "magic" in the sense that there is no empirical or logical basis for assuming, in absence of evidence to the contrary, that a particular case is a average member of a set of similar cases.

> "To transform the absence of a reason into a positive reason represents a feat of oratorical art that is worthy of an attorney for the defense but is not permissible in the court of logic." [74, p. 354]

Trying to justify randomization empirically or logically is futile. Instead, we can view it as inducing a falsifiable scientific theory upon which to provisionally base our beliefs.

> "So it is a profound mistake to try to do what scientists and philosophers have almost always tried to do, namely prove the truth of a theory, or justify our belief in a theory, since this is to attempt the logically impossible. What we can do, however, and this is of the highest possible importance, is to justify our preference for one theory over another. If we are rational we shall always base our decisions and expectations on 'the best of our knowledge' ...and provisionally assume the 'truth' of that knowledge for practical purposes, because it is the least insecure foundation available; but we shall never lose sight of the fact that at any time experience may show it to be wrong and require us to revise it." [44, p. 26-27]

The implication is that, while our proposed reference class selection policy cannot be justified empirically or logically, we can justify the preference for it over other policies. The merit of our reference class selection policy is indicated by comparing the empirical results in section 6.4 with those of its competitors.

Our reference class selection policy is not without its shortcomings however. As will be seen, the policy is related to the maximum entropy principle [28], and shares the problem of computation—there is no known way to compute it in general. While this is inconvenient for practical purposes, the policy is nevertheless useful as a specification of the desired properties of our inferences and can be computed in restricted settings.

In the next chapter, the reference class selection policy developed in this chapter is used in the specification of defaults in Meta-Theorist.

5.1 Problems Inheriting Statistics

> "The fact that many properties in the most specific reference class are irrelevant is the intuitive basis for allowing the inheritance of expected statistics, and the default assumptions enable this kind of inheritance. The preference criterion captures the natural constraint that we should try to retain as much information as possible; i.e., we should use the narrowest reference class possible." [3, p. 167]

In Bacchus's reference class selection policy, assumed irrelevance allows moving to a wider reference class while specificity (known relevance) constrains the widening. For instance, the assumed irrelevance[3] of blackness in Example 4.2 allowed moving from T_0 to T_1 while the specificity of the antarctic bird statistics in comparison to the bird statistics disallowed moving to T_2.

[3] The actual assumption is about "expectation independence" rather than statistical independence.

In addition to the specificity constraint, Definition 25 restricts inheritance of expected statistics to *downward inheritance* from known supersets. These two constraints are the key determinants of viable reference classes (with corresponding viable theories) in Bacchus's formalism. In this section, two examples illustrate that Bacchus's reference class selection policy fails to sanction certain intuitive theories and fails to prohibit certain overly presumptuous theories. This informal discussion is put on firmer footing in the next section.

Consider the Black Bird Problem in Example 5.1. Here we have that

SKB5.1:

$\texttt{bird(tweety)}$,
$\texttt{black(opus)} \;\&\; \texttt{bird(opus)}$,

And either (a):

$[\texttt{fly(X)}|\texttt{bird(X)}]_{\texttt{X}} = 0.75;$

or else (b):

$[\texttt{fly(X)}|\texttt{black(X)} \;\&\; \texttt{bird(X)}]_{\texttt{X}} = 0.75.$

Example 5.1: The Black Bird Problem

Tweety is a bird and Opus is a black bird. If, as in SKB5.1a, it was also believed that 75% of birds fly then there would be a viable theory containing

$$\mathsf{E}\Big([\texttt{fly(V)}|\texttt{black(V)} \;\&\; \texttt{bird(V)}]_{\texttt{V}}\Big) = \mathsf{E}\Big([\texttt{fly(V)}|\texttt{bird(V)}]_{\texttt{V}}\Big)$$

and this entails

$$\mathsf{prob}\Big(\mathtt{fly(opus)}\Big) = 0.75.$$

That is, the Definition 25 permits the *downward inheritance* of the expected statistics of flying for birds to black birds via the assumed expectation independence of blackness w.r.t. flying given bird. Yet, if instead, as in SKB5.1b, it was believed that 75% of black birds fly then there is no viable theory containing

$$\mathsf{E}\Big([\mathtt{fly(V)}|\mathtt{black(V)}\ \&\ \mathtt{bird(V)}]_{\mathsf{V}}\Big) = \mathsf{E}\Big([\mathtt{fly(V)}|\mathtt{bird(V)}]_{\mathsf{V}}\Big)$$

and so

$$\mathsf{prob}\Big(\mathtt{fly(tweety)}\Big) = 0.75$$

is not concluded. Definition 25 prohibits the *upward inheritance* of the expected statistics of flying for black birds to birds, i.e., the expectation independence of blackness w.r.t. flying given bird cannot be assumed. This is peculiar—if blackness is expectation independent of flying given bird then knowing the flying statistic for either bird or black bird determines the statistic for the other—there seems to be no intuitive justification for this prohibition of upward inheritance.

Failing to draw intuitive conclusions is a problem, but a more serious problem is drawing counterintuitive conclusions. Consider the Vaccinated Child Problem in Example 5.2. Here we have that vaccinated people probably

SKB5.2·

$$[\mathtt{immune(X)}|\mathtt{vaccinated(X)}]_{\mathtt{X}} = 0.75,$$
$$\forall \mathtt{X}.(\mathtt{blonde(X)}\ \&\ \mathtt{vaccinated\_child(X)}) \rightarrow \neg\mathtt{immune(X)},$$
$$\forall \mathtt{X}.\mathtt{vaccinated\_child(X)} \rightarrow \neg\mathtt{immune(X)},$$
$$\mathtt{vaccinated\_child(mary)}.$$

Example 5.2: The Vaccinated Child Problem

acquire immunity (0.75), blonde vaccinated children do not acquire immunity, and Mary is a vaccinated child.

By direct inference we have that T_0 contains

$$\mathsf{prob}\Big(\mathtt{immune(mary)}\Big) = \mathsf{E}\Big([\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big).$$

This tells us nothing useful since the reference class vaccinated child is overly specific (i.e., we have no statistic for immune given vaccinated child).

We could, however, infer

$$\mathsf{prob}\Big(\mathtt{immune(mary)}\Big) = 0.75$$

from the viable theory T_1 in which we assume

$$\mathsf{E}\Big([\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big) = \mathsf{E}\Big([\mathtt{immune(V)}|\mathtt{vaccinated(V)}]_{\mathtt{V}}\Big).$$

That is, by T_0, which is contained in T_1,

$$\mathsf{prob}\Big(\mathtt{immune(mary)}\Big) = \mathsf{E}\Big([\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big)$$

and by $T1$,

$$\mathsf{E}\Big([\mathtt{immune(V)|vaccinated\_child(V)}]_{\mathtt{V}}\Big) = \mathsf{E}\Big([\mathtt{immune(V)|vaccinated(V)}]_{\mathtt{V}}\Big)$$

and from the KB (and properties of E) we know

$$\mathsf{E}\Big([\mathtt{immune(V)|vaccinated(V)}]_{\mathtt{V}}\Big) = 0.75$$

so we have

$$\mathsf{prob}\Big(\mathtt{immune(mary)}\Big) = 0.75.$$

We also have that T_1 contains

$$\mathsf{prob}\Big(\mathtt{blonde(mary)}\Big) \leq 0.25$$

as shown in the proof sketch in Figure 14.

Notice that the expectation independence assumption has the consequence that

$$\mathsf{E}\Big([\mathtt{blonde(V)|vaccinated\_child(V)}]_{\mathtt{V}}\Big) \leq 0.25.$$

This constraint on the expected proportion of blonde vaccinated children among vaccinated children seems unwarranted—since most vaccinated people acquire immunity, vaccinated children are vaccinated people, and blonde vaccinated children do not acquire immunity implies *either* most vaccinated children are not blonde *or* most vaccinated people are not vaccinated children—as the proof sketch in Figure 15 shows. Making the expectation independence assumption has the effect of choosing the first of the two disjuncts.

$$\begin{aligned}
&\mathsf{E}\Big([\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big) \\
&\quad = \mathsf{E}\Big([\mathtt{immune(V)}|\mathtt{vaccinated(V)}]_{\mathtt{V}}\Big) && \text{Def. 25} \\
&\mathsf{E}\Big([\mathtt{immune(V)}|\mathtt{vaccinated(V)}]_{\mathtt{V}}\Big) = 0.75 \\
&\mathsf{E}\Big([\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big) = 0.75 \\
&\mathsf{E}\Big([\neg\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big) = 0.25 \\
&\mathsf{cert}\Big([\mathtt{blonde(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}} \\
&\quad \leq [\neg\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big) && \text{Lem. 16} \\
&\mathsf{E}\Big([\mathtt{blonde(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big) \\
&\quad \leq \mathsf{E}\Big([\neg\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big) && \text{Lem. 20.5} \\
&\mathsf{E}\Big([\mathtt{blonde(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big) \leq 0.25 \\
&\mathsf{prob}\Big(\mathtt{blonde(mary)}\Big) \\
&\quad = \mathsf{E}\Big([\mathtt{blonde(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big) && \text{Def. 22} \\
&\mathsf{prob}\Big(\mathtt{blonde(mary)}\Big) \leq 0.25
\end{aligned}$$

Figure 14: $\mathsf{prob}\Big(\mathtt{blonde(mary)}\Big) \leq 0.25$

$$[\texttt{immune(V)}|\texttt{vaccinated(V)}]_{\texttt{V}} = 0.75$$

$$\forall\texttt{V.}\big(\texttt{blonde(V)}\ \&\ \texttt{vaccinated\_child(V)}\big) \rightarrow \neg\texttt{immune(V)}$$

$$\begin{aligned}&[\neg(\texttt{blonde(V)}\ \&\ \texttt{vaccinated\_child(V)})|\texttt{vaccinated(V)}]_{\texttt{V}}\\ &\quad \geq [\texttt{immune(V)}|\texttt{vaccinated(V)}]_{\texttt{V}} = 0.75 \qquad \text{Lem. 16}\end{aligned}$$

$$\begin{aligned}&[\texttt{blonde(V)}\ \&\ \texttt{vaccinated\_child(V)}|\texttt{vaccinated(V)}]_{\texttt{V}}\\ &\quad \leq 0.25\end{aligned}$$

$$\begin{aligned}&[\texttt{blonde(V)}|\texttt{vaccinated\_child(V)}\ \&\ \texttt{vaccinated(V)}]_{\texttt{V}}\\ &\quad \times[\texttt{vaccinated\_child(V)}|\texttt{vaccinated(V)}]_{\texttt{V}} \leq 0.25\end{aligned}$$

$$\begin{aligned}&[\texttt{blonde(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}\\ &\quad \times[\texttt{vaccinated\_child(V)}|\texttt{vaccinated(V)}]_{\texttt{V}} \leq 0.25 \qquad \text{Lem. 15}\end{aligned}$$

$$\begin{aligned}&[\texttt{blonde(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}} \leq 0.25\\ &\quad \texttt{or}\ [\texttt{vaccinated\_child(V)}|\texttt{vaccinated(V)}]_{\texttt{V}} \leq 0.25\end{aligned}$$

Figure 15: Is $[\texttt{blonde(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}} \leq 0.25$?

Another possibility is that the expected proportion with immunity among vaccinated children is the same as the expected proportion with immunity among blonde vaccinated children:

$$\begin{aligned}
&\mathsf{E}\Big([\texttt{immune(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big)\\
&= \mathsf{E}\Big([\texttt{immune(V)}|\texttt{blonde(V)}\ \&\ \texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big).
\end{aligned}$$

If this is the case, then the expected proportion of blonde vaccinated children is unconstrained (Figure 16) while the expected proportion of vaccinated children among vaccinated people is constrained (Figure 17).

$$\begin{array}{ll}
\mathsf{cert}\Big([\neg\texttt{blonde(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}} & \\
\quad \geq [\texttt{immune(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big) & \text{Lem. 16} \\
\mathsf{E}\Big([\neg\texttt{blonde(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big) & \\
\quad \geq \mathsf{E}\Big([\texttt{immune(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big) & \text{Lem. 20.6} \\
\mathsf{E}\Big([\texttt{immune(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big) & \\
\quad = \mathsf{E}\Big([\texttt{immune(V)}|\texttt{blonde(V)}\ \&\ \texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big) & \mathsf{E}\text{-assum.} \\
\mathsf{E}\Big([\neg\texttt{blonde(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big) & \\
\quad \geq \mathsf{E}\Big([\texttt{immune(V)}|\texttt{blonde(V)}\ \&\ \texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big) & \\
\quad = \texttt{0} & \\
\mathsf{E}\Big([\texttt{blonde(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big) \leq \texttt{1} &
\end{array}$$

Figure 16: $\mathsf{E}\Big([\texttt{blonde(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big) \leq \texttt{1}$

Though there seems to be no intuitive reason to prefer one argument over

$$\mathtt{E}\Big([\mathtt{immune(V)}|\mathtt{blonde(V)}\ \&\ \mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big) = 0$$

$$\begin{aligned}&\mathtt{E}\Big([\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big)\\ &\quad = \mathtt{E}\Big([\mathtt{immune(V)}|\mathtt{blonde(V)}\ \&\ \mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big) \quad \text{E-assum.}\end{aligned}$$

$$\mathtt{E}\Big([\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big) = 0$$

$$\mathtt{cert}\Big([\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}} = 0\Big) \qquad \text{Lem. 20.7}$$

$$\begin{aligned}&\mathtt{cert}\Big([\neg\mathtt{vaccinated\_child(V)}|\mathtt{vaccinated(V)}]_{\mathtt{V}}\\ &\quad \geq [\mathtt{immune(V)}|\mathtt{vaccinated(V)}]_{\mathtt{V}} = 0.75\Big) \qquad \text{Lem. 16}\end{aligned}$$

$$\mathtt{E}\Big([\mathtt{vaccinated\_child(V)}|\mathtt{vaccinated(V)}]_{\mathtt{V}}\Big) \leq 0.25$$

Figure 17: $\mathtt{E}\Big([\mathtt{vaccinated\_child(V)}|\mathtt{vaccinated(V)}]_{\mathtt{V}}\Big) \leq 0.25$

the other, the latter argument is ruled out by Definition 25 which prohibits upward inheritance of expected statistics from subsets to supersets.

An important objection to the analysis presented in this section is that implicit linguistic information has not been considered. For instance, one might claim that vaccinated children should inherit statistics from vaccinated people on the grounds that if vaccinated children were exceptional vaccinated people with respect to immunity then we should have reported their statistics. That is, what we do not *say* is as iinportant as what we do say. This objection, however, can be safely dismissed because the statistical knowledge bases are intended to represent what is fully believed by the agent. If there are any linguistic conventions in effect, they should be part of the knowledge base. Yet we should bear this objection in mind when we come across counterintuitive examples—the counterintuitiveness might be explained by our own hidden linguistic conventions.

The two examples in this section illustrate that Bacchus's reference class selection policy does not take upward inheritance into account. This results in irrelevant information blocking legitimate inheritance as in Example 5.1 or relevant information failing to block illegitimate (overly presumptuous) inheritance[4] as in Example 5.2. To replace the inadequate Definition 25 in Bacchus's formalism, the next section develops a mathematical basis for expectation independence assumptions which shows the legitimacy of upward inheritance.

5.2 Average Subsets

The concept of the average subset is important because it is general set-theoretic fact that the mean of all subsets of a set equals the mean of the set (cf. [82, 33]). Inheritance between a set and a subset can be considered legitimate if the subset is indistinguishable from the average subset (at least w.r.t. the inherited property). This is because the expectation independence

[4] More precisely, Bacchus views both T_0 and T_1 from Example 5.1 as viable theories and does not choose between them. Additionally, Bacchus does not view the theory T_2 in which we assume $\mathsf{E}([\texttt{immune(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}) = \mathsf{E}([\texttt{immune(V)}|\texttt{blonde(V) \& vaccinated\_child(V)}]_{\texttt{V}})$ as viable. Consequently, upward inheritance is not considered viable while both no inheritance and downward inheritance are considered viable.

assumptions, which sanction inheritance, hold between a set and the average subset. Consequently, the test for viable inheritance reduces to a test for indistinguishability of a subset of a set and the average subset w.r.t. the inherited property. Bear in mind, however, that the average subset is just an abstraction and it may be the case that there is no subset with average characteristics.

To employ the concept of average subset in testing for viable inheritance in Bacchus's formalism, a distinction must first be drawn between expectation over possible worlds and expectation over the possible interpretations of predicates.

Definition 31 (WORLD EXPECTATION (BACCHUS))
The *world expectation* operator E applied to its operand denotes the weighted average of the operand across the possible worlds where the weightings are determined by the probability distribution μ_S over possible worlds (cf. [3, p. 135]).

Bacchus's world expectation operator applies over possible worlds. These worlds are interpretations of predicates, terms, and constants. We now wish to introduce a more restrictive expectation operator which applies over interpretations of predicates (with a fixed interpretation of terms and constants) and which weights each interpretation equally. A formal extension to Bacchus's probability logic is necessary to include this predicate interpretation expectation, but we do not intend to pursue this here.

Definition 32 (PREDICATE INTERPRETATION EXPECTATION)
Let the predicates of a given KB be divided (as specified in advance) into two classes: *fixed* and *varying.* Then within a given possible world, the *predicate interpretation expectation* operator $\mathcal{E}$ applied to its operand denotes the (unweighted) average of the operand across every interpretation which satisfies KB and which has the interpretation of the fixed predicates, terms and constants fixed in advance (as determined by the possible world within which $\mathcal{E}$ is applied).

We indicate varying predicates by a subscript and refer to them as predicate variables. We refer to fixed predicates as predicate constants. For example, the expression $\mathcal{E}\Big([\alpha|\beta]_{\mathtt{V}}\Big)_{\beta}$ means the average of $[\alpha|\beta]_{\mathtt{V}}$ across the possible interpretations of β. (For the expectation to make sense, it must be the case that $[\beta]_{\mathtt{V}} > 0$.)

As a concrete example, suppose we have a domain of only two birds: `heckle` and `jeckle`. Suppose that `heckle` flies and `jeckle` does not. Suppose there is one further individual who is not a bird: `fred`. From this we have that $[\mathtt{fly(X)}|\mathtt{bird(X)}]_{\mathtt{X}} = 0.5$. In evaluating

$$\mathcal{E}\Big([\mathtt{fly(V)}|\mathtt{black(V)}\ \&\ \mathtt{bird(V)}]_{\mathtt{V}}\Big)_{\mathtt{<black>}}$$

that is, in determining the average value of $[\mathtt{fly(V)}|\mathtt{black(V)}\ \&\ \mathtt{bird(V)}]_{\mathtt{V}}$ over varying interpretations of `black`, we let the interpretation of `black` range over subsets of the domain that satisfy the KB. For the expected value to be defined, we require that [black(V) & bird(V)]v > 0. There are six subsets of the domain that satisfy the constraints. This results in three possible sets of black birds: { heckle, jeckle }, {heckle }, and {jeckle}. The value of [fly(V)|black(V) & bird(V)]v for each of these interpretations is: 0.5, 1, and 0 respectively. The expected value is the (unweighted) average of these: 0.5.

The following theorem is a restatement of the general set-theoretic fact that the mean of all subsets of a set equals the mean of the set (cf. [82, 33]).

Theorem 33 (Averaged Predicates)

Let β be a predicate variable and let α and γ be predicaɩe constants. If the interpretation of β ranges over every subset of γ then

$$\mathcal{E}\left([\alpha|\beta]_{\mathsf{V}}\right)_{\beta} = [\alpha|\gamma]_{\mathsf{V}}.$$

This theorem brings us close to the (world) expectation independence assumptions we are seeking as seen in the following theorem.

Theorem 34 (World Expectation for Averaged Predicates)

Let β be a predicate variable and let α and γ be predicate constants. If the interpretation of β ranges over every subset of γ then

$$\mathsf{E}\left(\mathcal{E}\left([\alpha|\beta]_{\mathsf{V}}\right)_{\beta}\right) = \mathsf{E}\left(\mathcal{E}\left([\alpha|\gamma]_{\mathsf{V}}\right)_{\beta}\right).$$

Proof: Since $[\alpha|\gamma]_V$ is constant within each world (only the interpretation of β is varied), $\mathcal{E}\left([\alpha|\gamma]_{\mathtt{V}}\right)_\beta = [\alpha|\gamma]_{\mathtt{V}}$ within each world. And since β satisfies the condition of Theorem 33 (in every world), $\mathcal{E}\left([\alpha|\beta]_{\mathtt{V}}\right)_\beta = [\alpha|\gamma]_{\mathtt{V}}$ in every world. Consequently, $\mathsf{cert}\left(\mathcal{E}\left([\alpha|\beta]_{\mathtt{V}}\right)_\beta = \mathcal{E}\left([\alpha|\gamma]_{\mathtt{V}}\right)_\beta\right)$. Hence by Lemma 20.2, $\mathsf{E}\left(\mathcal{E}\left([\alpha|\beta]_{\mathtt{V}}\right)_\beta\right) = \mathsf{E}\left(\mathcal{E}\left([\alpha|\gamma]_{\mathtt{V}}\right)_\beta\right)$. ■

This provides a mathematical basis for upward and downward inheritance of expected statistics. The expected statistics of a set and the average subset of the set are equivalent. Therefore, if one set is an average subset of another, we can inherit expected statistics from one set to the other in either direction.

In Example 5.1 from the previous section, if black bird is an average subset of bird then

$$\mathsf{E}\left([\mathtt{fly(V)}|\mathtt{black(V)\ \&\ bird(V)}]_{\mathtt{V}}\right)_{\mathtt{<black(V)\ \&\ bird(V)>}} = \mathsf{E}\left([\mathtt{fly(V)}|\mathtt{bird(V)}]_{\mathtt{V}}\right)$$

which entails

$$\mathsf{prob}\left(\mathtt{fly(tweety)}\right) = \mathsf{prob}\left(\mathtt{fly(opus)}\right).$$

That is, the degree of belief in flying is the same for birds and black birds.

In Example 5.2, if vaccinated children is an average subset of vaccinated people then

$$\begin{aligned}&\mathsf{E}\left([\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\right)\\ &= \mathsf{E}\left([\mathtt{immune(V)}|\mathtt{vaccinated(V)}]_{\mathtt{V}}\right)\end{aligned}$$

and if blonde vaccinated children is an average subset of vaccinated children then

$$\begin{aligned}&\mathsf{E}\Big([\mathtt{immune(V)}|\mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big)\\&= \mathsf{E}\Big([\mathtt{immune(V)}|\mathtt{blonde(V)}\ \&\ \mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big).\end{aligned}$$

It is clear that both cannot be the case since SKB 5.2 implies

$$\begin{aligned}&[\mathtt{immune(V)}|\mathtt{vaccinated(V)}]_{\mathtt{V}} = 0.75 \text{ and}\\&[\mathtt{immune(V)}|\mathtt{blonde(V)}\ \&\ \mathtt{vaccinated\_child(V)}]_{\mathtt{V}} = 0\end{aligned}$$

which implies

$$\begin{aligned}&\mathsf{E}\Big([\mathtt{immune(V)}|\mathtt{vaccinated(V)}]_{\mathtt{V}}\Big)\\&\neq \mathsf{E}\Big([\mathtt{immune(V)}|\mathtt{blonde(V)}\ \&\ \mathtt{vaccinated\_child(V)}]_{\mathtt{V}}\Big).\end{aligned}$$

Although we are mathematically justified in inheriting expected statistics between a set and an average subset, we generally do not know whether the subsets in the knowledge base are average. We do not know, for instance, that vaccinated children is an average subset of vaccinated people. In fact, it seems pretty obvious that the average vaccinated person is not a child. While the average subset is representative of the set, we do not know whether the particular subset we happen to have is representative. This problem is analogous to the problem of the single case probability. For that problem, the direct inference principle (Definition 22) uses the technique of randomization. The probability for a particular case is determined by the probability average case. The solution to our current problem is likewise the technique of randomization—in this case, a second order randomization.

5.3 Second Order Randomization

> "The probability of an event is determined by our own state of knowledge and ignorance. We need knowledge of the relevant measure statements; we may ignore *special characteristics* of the object or event under consideration which are not *known* to be related to the property in question." [33, p. 185]

The inheritance of statistics (in either direction) between a superset and a subset hinges on a *second order randomization* assumption, i.e., the assumption that as far as we know, the subset in question is an average subset of the superset w.r.t. the inherited property.

The following definitions are only informal as they involve extending Bacchus's probability logic to include *predicate variables.*

Definition 35 (SECOND ORDER RANDOMIZATION)

Let α be a formula of $\mathcal{L}^{\mathbf{stat}}$. If $\langle p_1, \ldots, p_n \rangle$ are n distinct predicate symbols appearing in α and $\langle pv_1, \ldots, pv_n \rangle$ are n distinct predicate variables that do not occur in α, then let $\alpha^{\mathbf{pv}}$ denote the new formula which results from textually substituting p_i by pv_i in α, for all i.

A preliminary Second Order Direct Inference Principle is stated in Definition 36. We first examine the properties of this definition and then discuss a more elaborate version (Definition 37).

Definition 36 (Prelim. Second Order Direct Inference Principle)

If α is a formula of $\mathcal{L}^{\mathsf{stat}}$ and if KB is the complete set of objective formulas that the agent fully believes, then the agent's degrees of belief should be determined[5] by the equality

$$\mathsf{prob}(\alpha) = \mathsf{E}\Big(\mathcal{E}\big([\alpha^{\mathsf{pv}}|(\mathsf{KB}^{\mathsf{v}})^{\mathsf{pv}}]_{\mathsf{v}}\big)_{\mathsf{pv}}\Big)$$

where the predicate variables $\langle pv_1, \ldots, pv_n\rangle$ range over predicate interpretations satisfying $(\mathsf{KB}^{\mathsf{v}})^{\mathsf{pv}}$. The agent must also fully believe that $\mathcal{E}\big([(\mathsf{KB}^{\mathsf{v}})^{\mathsf{pv}}]_{\mathsf{v}}\big) > 0$, i.e., $\mathsf{cert}\Big(\mathcal{E}\big([(\mathsf{KB}^{\mathsf{v}})^{\mathsf{pv}}]_{\mathsf{v}}\big) > 0\Big)$.

Recall that the (first order) direct inference principle (Definition 22) constrains degrees of belief by replacing each constant of KB with a random designator. The result is that the properties of particular individuals are determined by the properties of the set of similar individuals. In the same fashion, this preliminary second order direct inference principle further constrains degrees of belief by replacing each predicate of KB by a predicate variable that ranges over interpretations satisfying the KB. The result in this case is that the (world) expectation for particular predicates is determined by the (predicate interpretation) expectation for the set of similar predicates.

For instance, suppose all we know is that most vaccinated people acquire

[5] This definition subsumes First Order Direct Inference.

immunity, and vaccinated children are vaccinated people, and Mary is a vaccinated child as in Example 5.3. We can use (preliminary) second order direct inference to determine the degree of belief in Mary acquiring immunity (for $\mathcal{R}$ = <immune,vaccinated,vaccinated_child>).

SKB5.3:

$[\texttt{immune(X)}|\texttt{vaccinated(X)}]_{\texttt{X}} = 0.75,$
$\forall\texttt{X.vaccinated\_child(X)} \rightarrow \texttt{vaccinated(X)}.$
$\texttt{vaccinated\_child(mary)}.$

Example 5.3: Second Order Randomization

$$
\begin{aligned}
&\texttt{prob(immune(mary))}\\
&\quad = \mathsf{E}\Big(\mathcal{E}\Big([\texttt{immune(V)}|\texttt{vaccinated\_child(V)}\ \&\\
&\qquad\quad [\texttt{immune(X)}|\texttt{vaccinated(X)}]_{\texttt{X}} = 0.75\ \&\\
&\qquad\quad \forall\texttt{X.vaccinated\_child(X)} \rightarrow \texttt{vaccinated(X)}]_{\texttt{V}}\Big)_{\mathcal{R}}\Big) && \text{Def. 37}\\
&\quad = \mathsf{E}\Big(\mathcal{E}\Big([\texttt{immune(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big)_{\mathcal{R}}\Big) && \text{Cor. 13}\\
&\quad = \mathsf{E}\Big(\mathcal{E}\Big([\texttt{immune(V)}|\texttt{vaccinated(V)}]_{\texttt{V}}\Big)_{\mathcal{R}}\Big) && \text{Thm. 34}\\
&\quad = \mathsf{E}\Big([\texttt{immune(V)}|\texttt{vaccinated(V)}]_{\texttt{V}}\Big)\\
&\quad = 0.75
\end{aligned}
$$

The argument relies on an indirect use of Theorem 34. Once we pick the interpretation of the predicates **immune** and **vaccinated**, since the choice must satisfy $(\texttt{KB}^{\texttt{V}})^{\texttt{PV}}$, we have both that $[\texttt{immune(X)}|\texttt{vaccinated(X)}]_{\texttt{X}} = 0.75$

and that the interpretation of the predicate vaccinated_child must range over subsets of vaccinated, i.e., it is an average subset of vaccinated.

The following (world) expectation independence

$$\mathsf{E}([\alpha|\beta]_{\mathsf{v}}) = \mathsf{E}([\alpha|\gamma]_{\mathsf{v}})$$

results[6] from the second order direct inference principle combined with Theorem 34 provided KB is such that

$$\mathsf{E}(\mathcal{E}([\alpha^{\mathtt{pv}}|\beta^{\mathtt{pv}} \wedge (\mathtt{KB}^{\mathtt{v}})^{\mathtt{pv}}]_{\mathsf{v}})_{\mathtt{pv}}) = \mathsf{E}(\mathcal{E}([\alpha^{\mathtt{pv}}|\gamma^{\mathtt{pv}} \wedge (\mathtt{KB}^{\mathtt{v}})^{\mathtt{pv}}]_{\mathsf{v}})_{\mathtt{pv}}).$$

If this condition is satisfied (which means roughly that on average the differences between β and γ do not cause their α-statistic to differ) then inheritance between β and γ is viable. However, KB might contain information that interferes with this inference.

For instance in Example 5.3, had we known the immunity statistic for blonde vaccinated children and vaccinated people differ (as in Example 5.4), the interpretation of the predicate vaccinated_child would be constrained in a way that prohibits it from being an average subset of vaccinated. This can be shown by computing the statistics over the space of possible interpretations for a given universe of discourse. In Example 5.4, if the universe of discourse consists of four individuals, we can compute by exhaustive enumeration that (for $\mathcal{R}$ = <immune,vaccinated,vaccinated_child,blonde>)

[6] Provided $\mathsf{E}([\alpha|\beta]_{\mathsf{v}}) = \mathsf{E}([\alpha|\beta \wedge \mathtt{KB}^{\mathtt{v}}]_{\mathsf{v}})$ and $\mathsf{E}([\alpha|\gamma]_{\mathsf{v}}) = \mathsf{E}([\alpha|\gamma \wedge \mathtt{KB}^{\mathtt{v}}]_{\mathsf{v}})$.

$$\mathsf{E}\Big(\mathcal{E}\Big([\texttt{immune(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}\Big)_{\mathcal{R}}\Big) = 0.625$$
$$\neq \mathsf{E}\Big(\mathcal{E}\Big([\texttt{immune(V)}|\texttt{vaccinated(V)}]_{\texttt{V}}\Big)_{\mathcal{R}}\Big) = 0.75.$$

So in this case, inheritance of the immunity statistics from vaccinated people to vaccinated_child is blocked.

SKB5.4:

$[\texttt{immune(X)}|\texttt{vaccinated(X)}]_{\texttt{X}} = 0.75,$
$\forall \texttt{X.vaccinated\_child(X)} \rightarrow \texttt{vaccinated(X)},$
$[\texttt{immune(X)}|\texttt{blonde(X) \& vaccinated\_child(X)}]_{\texttt{X}} = 0.5,$
$\texttt{vaccinated(mary)}.$

Example 5.4: Interference

While the preliminary second order direct inference principle results in inheritance that is more in line with our intuitions than that resulting from Bacchus's nonmonotonic assumptions, there are at least two peculiarities.

First, every proposition is assigned a degree of belief. Propositions for which we have no relevant knowledge take on mid-value degrees of belief (e.g., prob(tall(mary)) = 0.5). Propositions for which there are conflicting relevant statistics take on mid-value degrees of belief (e.g., prob(immune(mary)) takes on the mid-value between the statistic for immune vaccinated people and the statistic for blonde vaccinated children). This is perhaps too committal.

Second, the results are syntax sensitive. In Example 5.5, which is a logically equivalent axiomatization of Example 5.4, we have (again for a universe of discourse with four individuals)

$$\begin{aligned}&\mathsf{E}\Big(\mathcal{E}\big([\texttt{immune(V)}|\texttt{vaccinated\_child(V)}]_{\texttt{V}}\big)_{\mathcal{R}}\Big) \approx 0.667\\ &\neq \mathsf{E}\Big(\mathcal{E}\big([\texttt{immune(V)}|\texttt{vaccinated(V)}]_{\texttt{V}}\big)_{\mathcal{R}}\Big) = 0.75\end{aligned}$$

(for $\mathcal{R}$ = <immune,vaccinated,vaccinated_child,blonde_vaccinated_child>). Comparing this with the result for Example 5.4 shows that the expected statistics are syntax sensitive (i.e., the proportion of immune vaccinated children is 0.625 in one case and (approx.) 0.667 in the other). More precisely, the difference comes about because the predicate variables differ. In one case, we have blonde as a predicate variable and in the other case we have blonde_vaccinated_child. This raises an interesting question about what to designate as the predicate variables. A similar issues arises in circumscription [47], namely, which predicates should be allowed to vary and which should be fixed. We will not pursue these issues here.

SKB5.5:

$[\texttt{immune(X)}|\texttt{vaccinated(X)}]_{\texttt{X}} = 0.75$,
$\forall\texttt{X.vaccinated\_child(X)} \rightarrow \texttt{vaccinated(X)}$,
$\forall\texttt{X.blonde\_vaccinated\_child(X)} \rightarrow \texttt{vaccinated\_child(X)}$,
$[\texttt{immune(X)}|\texttt{blonde\_vaccinated\_child(X)}]_{\texttt{X}} = 0.5$,
$\texttt{vaccinated\_child(mary)}$.

Example 5.5: Syntax Sensitive Expectation

Bacchus and Halpern are investigating the properties of this preliminary second order direct inference principle[7] and it appears to be connected to the maximum entropy principle [28]. They are taking a slightly different approach that essentially flattens out the $\mathsf{E}(\mathcal{E}(-))$ term by placing a uniform distribution over possible worlds and ranging them over possible interpretations of the predicates. This combines the two expectations.

Earlier in this section, we noted a condition for viable inheritance. We can use this to avoid the peculiarities of the preliminary second order direct inference principle as follows.

Definition 37 (SECOND ORDER DIRECT INFERENCE PRINCIPLE)

If α, β and γ are formulas of $\mathcal{L}^{\mathrm{stat}}$ and if KB is the complete set of objective formulas that the agent fully believes, then the agent's degrees of belief should be constrained[8] by the equality

$$\mathsf{E}([\alpha|\beta \wedge \mathsf{KB}^{\mathsf{v}}]_{\mathsf{v}}) = \mathsf{E}([\alpha|\gamma \wedge \mathsf{KB}^{\mathsf{v}}]_{\mathsf{v}})$$

if

$$\mathsf{E}(\mathcal{E}([\alpha^{\mathsf{pv}}|\beta^{\mathsf{pv}} \wedge (\mathsf{KB}^{\mathsf{v}})^{\mathsf{pv}}]_{\mathsf{v}})_{\mathsf{pv}}) = \mathsf{E}(\mathcal{E}([\alpha^{\mathsf{pv}}|\gamma^{\mathsf{pv}} \wedge (\mathsf{KB}^{\mathsf{v}})^{\mathsf{pv}}]_{\mathsf{v}})_{\mathsf{pv}})$$

where the predicate variables $\langle pv_1, \ldots, pv_n\rangle$ range over interpretations satisfying $(\mathsf{KB}^{\mathsf{v}})^{\mathsf{pv}}$. The agent must also fully believe that $\mathcal{E}([\beta^{\mathsf{pv}} \wedge (\mathsf{KB}^{\mathsf{v}})^{\mathsf{pv}}]_{\mathsf{v}})_{\mathsf{pv}} > 0$

[7] Personal communication.

[8] This Second Order Direct Inference Principle is applied in conjunction with (rather than in place of) the First Order Direct Inference Principle.

and $\mathcal{E}([\gamma^{\mathtt{pv}} \wedge (\mathtt{KB}^{\mathtt{v}})^{\mathtt{pv}}]_{\mathtt{v}})_{\mathtt{pv}} > 0$, i.e., $\mathtt{cert}(\mathcal{E}([\beta^{\mathtt{pv}} \wedge (\mathtt{KB}^{\mathtt{v}})^{\mathtt{pv}}]_{\mathtt{v}})_{\mathtt{pv}} > 0)$ and $\mathtt{cert}(\mathcal{E}([\gamma^{\mathtt{pv}} \wedge (\mathtt{KB}^{\mathtt{v}})^{\mathtt{pv}}]_{\mathtt{v}})_{\mathtt{pv}} > 0)$.

In the preliminary second order direct inference principle, (world) expectation for particular predicates is determined by (predicate interpretation) expectation for the set of similar predicates. In contrast, here (world) expectation *independence* for particular predicates is determined by (predicate interpretation) expectation *independence* for the set of similar predicates. This new definition does not result in every proposition being assigned a degree of belief and it appears not to be syntax sensitive. Note we can show the former by construction and, though we have no general proof of the latter, we can show it is true by exhaustive enumeration for particular examples like the one in connection with Example 5.3 and 5.4.

To illustrate second order direct inference, consider again Example 5.3. We can determine the degree of belief in Mary acquiring immunity as follows: (for $\mathcal{R}$ = <immune,vaccinated,vaccinated_child>)

$$\begin{aligned}
&\mathtt{prob(immune(mary))}\\
&= \mathtt{E([immune(V)|[immune(X)|vaccinated(X)]_X = 0.75\ \&}\\
&\qquad \mathtt{\forall X.vaccinated\_child(X) \rightarrow vaccinated(X)\ \&}\\
&\qquad \mathtt{vaccinated\_child(V)]_V)} \qquad\qquad \text{Def. 22}\\
&= \mathtt{E([immune(V)|[immune(X)|vaccinated(X)]_X = 0.75\ \&}\\
&\qquad \mathtt{\forall X.vaccinated\_child(X) \rightarrow vaccinated(X)\ \&}
\end{aligned}$$

$$\texttt{vaccinated(V)]}_{\texttt{V}}\texttt{)} \qquad \text{Def. 37}$$

$$= \texttt{E([immune(V)|vaccinated(V)]}_{\texttt{V}}\texttt{)} = 0.75$$

since

$$\texttt{E(}\mathcal{E}\texttt{([immune(V)|[immune(X)|vaccinated(X)]}_{\texttt{X}} = 0.75\ \&$$

$$\forall\texttt{X.vaccinated\_child(X)} \rightarrow \texttt{vaccinated(X)}\ \&$$

$$\texttt{vaccinated\_child(V)]}_{\texttt{V}}\texttt{)}_{\mathcal{R}}\texttt{)}$$

$$= \texttt{E(}\mathcal{E}\texttt{([immune(V)|vaccinated\_child(V)]}_{\texttt{V}}\texttt{)}_{\mathcal{R}}\texttt{)}$$

$$= \texttt{E(}\mathcal{E}\texttt{([immune(V)|vaccinated(V)]}_{\texttt{V}}\texttt{)}_{\mathcal{R}}\texttt{)} \qquad \text{Thm. 34}$$

$$= \texttt{E(}\mathcal{E}\texttt{([immune(V)|[immune(X)|vaccinated(X)]}_{\texttt{X}} = 0.75\ \&$$

$$\forall\texttt{X.vaccinated\_child(X)} \rightarrow \texttt{vaccinated(X)}\ \&$$

$$\texttt{vaccinated(V)]}_{\texttt{V}}\texttt{)}_{\mathcal{R}}\texttt{)}$$

Preliminary results suggest that Second Order Direct Inference achieves the desired inheritance for the examples commonly cited in the literature. For instance, simple inheritance, multiple inheritance with specificity, ambiguity, cascaded ambiguity, cycles, redundant information, negative paths, etc. all appear to be appropriately solved. This is supported in two ways. First, by the method of exhaustive enumeration for small examples (around four individuals and four predicates) and, second, by the Meta-Theorist implementation in the next chapter. Unfortunately, neither of these methods constitutes a proof in general.

Second Order Randomization also appears to exhibit the desirable property of weak chaining. Weak chaining involves inference through multiple probable (or default) links as seen in the following example. Suppose we know Tweety is a bird, 90% of birds are flying things, and 90% of flying things have wings. By Second Order Direct Inference, we can conclude at least 81% of birds have wings so the degree of belief in Tweety having wings is at least 0.81. This reasoning hinges on the conditional independence of wings and birds given flying—this independence is not given in the knowledge base but results from Second Order Randomization—i.e., it is essentially just an assumption. This form of reasoning may, of course, go wrong as can be seen by replacing "bird" in the above example by "helicopter." Nevertheless, weak chaining is generally a desirable property. Further work is necessary to prove the above properties hold in general.

Second Order Direct Inference can be determined mathematically, but we know of no easy way to do this in general. In the next chapter, we take the approach of identifying some characteristic patterns of interference to approximate the reference class selection policy based on Second Order Direct Inference described in this chapter.

Chapter 6

Statistically Motivated Defaults

This chapter provides a specification for defaults in Meta-Theorist. The key component of this specification is the theory pruning criteria for defaults. In addition to describing these criteria, the aim of this chapter is to provide an intuitive justification by appealing to probabilities and direct inference. The methodology employed is to compare a Meta-Theorist knowledge base with a corresponding statistical knowledge base. A mapping is also drawn between queries and between conclusions in the two representations. The pruning criteria are mapped to policies for choosing the reference class from the previous chapter. Because the pruning criteria are designed to make defaults applicable to a particular case when the corresponding statistical probability is applicable to that case, the defaults are said to be statistically motivated. Several examples are provided in this chapter to give some indication of the power and limitations of the proposed specification.

6.1 Interpreting Defaults, Queries, and Conclusions

The interpretation of defaults used here differs from the usual "assume true in the absence of evidence to the contrary" [72, 67]. Defaults are interpreted as asserting statistical knowledge which can be used to justify particular beliefs (cf. [3]). The following is the syntax for specifying defaults in Meta-Theorist.

default *Hypothesis*: $Wff_p \leftarrow (X) - Wff_c$. is synonymous with

hypothesis *Hypothesis*: $Wff_p \leftarrow Wff_c$.
meta fact *default(Hypothesis).*
meta fact *prediction(Hypothesis,* Wff_p*).*
meta fact *context(Hypothesis,* Wff_c*).*
meta fact *random(Hypothesis,X).*

The meta-predicate ***default*** distinguishes the hypothesis as a statistically motivated default, i.e., ***default(H)*** means $H \in \hat{\Delta}_{smd}$ where $\hat{\Delta}_{smd}$ is the subset of $\mathcal{H}$ containing hypotheses that are to be interpreted as statistically motivated defaults. The meta-predicate ***prediction*** identifies the prediction of the default—this plays the same role as the consequent (expression to the left of the conditioning bar) of a conditional probability term. The meta-predicate ***context*** identifies the context of the default—this plays the same role as the context (expression to the right of the conditioning bar) of a conditional probability term. Finally, the meta-predicate ***random*** indicates the variables to be interpreted as random designators. When all the variables of *Hypothesis* coincide with the random designators, the following abbreviated syntax is used.

default *Hypothesis*: $\mathit{Wff}_p \leftarrow \mathit{Wff}_c$.

Defaults are viewed as making statistical probability assertions about the domain. The statement

default birdsfly(X): fly(X) ← bird(X).

is interpreted as asserting that most birds fly, i.e., the proportion of birds that fly is greater than some constant (greater than 0.5). The same constant is assumed for all defaults and no commitment is made to its value (other than it being greater than 0.5). Consequently, defaults all have the same (unspecified) strength.

The above default can be understood as asserting

$$[\mathtt{birdsfly(X)} | \mathtt{bird(X)}]_{\mathtt{X}} > \mathtt{c}$$

$$\forall \mathtt{X}.(\mathtt{birdsfly(X)} \ \& \ \mathtt{bird(X)}) \rightarrow \mathtt{fly(X)}$$

which implies

$$[\mathtt{fly(X)} | \mathtt{bird(X)}]_{\mathtt{X}} > \mathtt{c}.$$

Definition 38 (INTERPRETATION OF DEFAULTS)

Defaults of the form

default H(X,Y): P(X,Y) ←(X)− C(X,Y).

correspond to the statistical assertions

$$\forall \mathtt{Y}.[\mathtt{H(X,Y)}|\mathtt{C(X,Y)}]_{\mathtt{X}} > \mathtt{c},$$

$$\forall \mathtt{X,Y}.(\mathtt{H(X,Y)}\ \&\ \mathtt{C(X,Y)}) \rightarrow \mathtt{P(X,Y)}.$$

In Example 6.1, the Meta-Theorist knowledge base MTKB6.1 corresponds to the statistical knowledge base SKB6.1.

MTKB6.1:

default birdsfly(X): fly(X) ← bird(X).
fact black(opus) & bird(opus).
conjecture blacksdontfly(X): ¬fly(X) ← black(X).

SKB6.1:

$[\mathtt{birdsfly(X)}|\mathtt{bird(X)}]_{\mathtt{X}} > \mathtt{c}$,
$\forall \mathtt{X}.(\mathtt{birdsfly(X)}\ \&\ \mathtt{bird(X)}) \rightarrow \mathtt{fly(X)}$,
$\mathtt{black(opus)}\ \&\ \mathtt{bird(opus)}$,
$\forall \mathtt{X}.(\mathtt{blacksdontfly(X)}\ \&\ \mathtt{black(X)}) \rightarrow \neg\mathtt{fly(X)}$

Example 6.1: Meta-Theorist KB and corresponding Statistical KB

The Meta-Theorist queries **explain** and **predict** are both interpreted as queries about propositional probabilities and each treats defaults in the same way. The difference is that explanation allows conjectures to be used while prediction does not.

The Meta-Theorist query (to MTKB6.1)

predict fly(opus).

can be informally understood as asking whether

$$\mathsf{prob}\big(\mathtt{fly(opus)}\big) > \mathsf{c}.$$

The answer to this query is understood as affirming that the above propositional probability can be inferred to be greater than c given SKB6.1 and the reference class selection policy. Note that the strength c of the answer to the query is the same as the strength of the defaults.

The Meta-Theorist query (to MTKB6.1)

explain fly(opus).

can be understood as asking whether, for some set of "acceptable" assumptions T,

$$\mathsf{prob}\big(\mathtt{fly(opus)}|\mathsf{T}\big) > \mathsf{c}.$$

For instance T might be {blacksdontfly(opus)}. The acceptability criterion for T depends on the kinds of hypotheses involved, e.g., for conjectures the criterion is consistency—see section 3.2.

6.2 Specifying Direct Inference

This section describes the specification of the Direct Inference Principle (Definition 22) by means of pruning criteria in Meta-Theorist. This results in a system that corresponds to T_0 from Chapter 4 for the restricted case where all known statistical probabilities are greater than some unspecified constant c, less than $1 - c$, or equal to one or zero.

The specification of the Direct Inference Principle is conceptually trivial. It involves simply determining the context of the query and whether the expression in the query either follows from the context plus the set of true assertions or whether the expression in the query follows from the prediction of some default whose context is equivalent to the context of the query. The former case covers direct inference from probability one statistical assertions (i.e., from the facts) and the latter covers direct inference from probability greater than c statistical assertions (i.e., from defaults). Direct inference amounts simply to instantiation of free variables (unification in Prolog) with the restriction that free variables in defaults are only instantiated if their context is equivalent to the context of the query.

The above relies on the following statistical facts:

1. if $[\alpha|\beta]_{\mathbf{x}} = 1$ (or 0) $\wedge\ \alpha \rightarrow \gamma$ then $[\alpha|\beta]_{\mathbf{x}} = [\gamma|\beta]_{\mathbf{x}}$;

2. if $\beta \equiv \gamma$ then $[\alpha|\beta]_{\mathbf{x}} = [\alpha|\gamma]_{\mathbf{x}}$; and

3. if $[\alpha|\beta]_{\mathbf{x}} > \mathrm{c} \wedge \alpha \rightarrow \gamma$ then $[\gamma|\beta]_{\mathbf{x}} \geq [\alpha|\beta]_{\mathbf{x}} > \mathrm{c}$.

As shown later in this section, we also need to make use of the following lemma which is an extension of Lemma 18.

Lemma 39

If no $x_i \in \vec{x}$ is free in the formula δ then

$$\begin{aligned}\models\ & \forall r_1 r_2.\Big[\beta \wedge \delta \wedge \forall \vec{z}(\alpha \rightarrow \gamma) \wedge [\alpha|\beta]_{\vec{x}} \in [r_1, r_2]\Big]_{\langle \vec{x}, \vec{y} \rangle} > 0 \\ & \rightarrow \Big[\gamma\Big|\beta \wedge \delta \wedge \forall \vec{z}(\alpha \rightarrow \gamma) \wedge [\alpha|\beta]_{\vec{x}} \in [r_1, r_2]\Big]_{\langle \vec{x}, \vec{y} \rangle} \in [r_1, 1].\end{aligned}$$

Proof: Follows from Lemma 16 and Lemma 18. ■

The main difficulty in specifying the Direct Inference Principle is determining the context of the query. Recall from Definition 22, that we determine the agent's degree of belief in α by

$$\mathsf{prob}(\alpha) = \mathsf{E}\Big([\alpha^{\mathsf{v}}|\mathsf{KB}^{\mathsf{v}}]_{\vec{v}}\Big).$$

In determining the context of the query, KB^{v} is simplified according to Lemma 12.

In Meta-Theorist, the corresponding context can be determined by a straightforward syntactic operation on the knowledge base. The meta-predicate *extract_context* extracts statements containing constants from the knowledge base and randomizes the constants (i.e., substitutes unique random variables).

The pruning criterion to specify direct inference (when considering only a single assumption) is:

```
meta fact prune([H]) ←
    default(H) &
    prediction(H,PH) &
    context(H,CH) &
```

```
        random(H,RVH) &
        extract_context(C0) &
        ¬equivalent_context(RVH,PH,CH,C0).
```

In words, direct inference using H (an instance of a default) is unacceptable if the context of the default and the context of the query are not equivalent.

Here *equivalent_context* is a meta-predicate that is true if the (randomized) context of H is equivalent to the context $C0$.

```
    meta fact equivalent_context(RVH,PH,CH,C0) ←
        simplify_context(RVH,PH,C0,Cn) &
        randomize(RVH,CH,CHV,[],VH) &
        equivalent(CHV,Cn).
```

Here *randomize* corresponds to randomization (Definition 21), i.e., the constants of CH are replaced by unique random variables (only the constants in RVH are randomized). The meta-predicate *equivalent* is true if its arguments are logically equivalent. Prior to checking for logical equivalence, the context $C0$ is simplified according to Lemma 39 by the meta-predicate *simplify_context*. The effect is to transform a query about whether

$$\left[\gamma\middle|\beta \wedge \delta \wedge \forall\vec{\mathbf{z}}((\alpha \wedge \beta) \rightarrow \gamma) \wedge [\alpha|\beta]_{\vec{\mathbf{x}}} > \mathsf{c}\right]_{\langle\vec{\mathbf{x}},\vec{\mathbf{y}}\rangle} > \mathsf{c}$$

to a query about whether

$$[\gamma|\beta]_{\vec{\mathbf{x}}} > \mathsf{c}.$$

The simplification algorithm is given in Figure 18. The value of C_i on the final iteration is the simplified context of the query.

```
simplify_context:

1. set $i = 0$;
2. set $C_0$ = the randomized facts, i.e., the facts of MTKB with constants
   replaced by unique variables;
3. set $V_i$ = the *random* variables of the randomized query;
4. repeat
    (a) increment $i$;
    (b) set $C_i$ = the statements of $C_0$ containing variables from $V_{i-1}$;
    (c) set $V_i$ = the variables of $C_i$;
   until $V_i = V_{i-1}$;
5. return *context* = $C_i$.
```

Figure 18: Algorithm to Simplify Meta-Theorist Context

The idea is this. When we are considering a query γ and there is a default α: $\gamma \leftarrow \beta$ and facts $\beta \wedge \delta$ then, in corresponding probabilistic terms, we are interested in

$$\left[\gamma \middle| \beta \wedge \delta \wedge \forall \vec{\mathbf{z}}((\alpha \wedge \beta) \rightarrow \gamma) \wedge [\alpha|\beta]_{\vec{\mathbf{x}}} > \mathsf{c}\right]_{\langle \vec{\mathbf{x}}, \vec{\mathbf{y}} \rangle} > \mathsf{c}.$$

A key observation is that whenever the conditions for Lemma 39 hold, we have

$$\begin{aligned} &\left[\gamma \middle| \beta \wedge \delta \wedge \forall \vec{\mathbf{z}}((\alpha \wedge \beta) \rightarrow \gamma) \wedge [\alpha|\beta]_{\vec{\mathbf{x}}} > \mathsf{c}\right]_{\langle \vec{\mathbf{x}}, \vec{\mathbf{y}} \rangle} \\ &= \left[\gamma \middle| \beta \wedge \delta \wedge \forall \vec{\mathbf{z}}((\alpha \wedge \beta) \rightarrow \gamma) \wedge [\alpha|\beta]_{\vec{\mathbf{x}}} > \mathsf{c}\right]_{\vec{\mathbf{x}}} \\ &= \left[\gamma \middle| \beta\right]_{\vec{\mathbf{x}}} \\ &\geq \left[\alpha \middle| \beta\right]_{\vec{\mathbf{x}}}. \end{aligned}$$

Therefore, in considering the query γ, it is sufficient to check whether the context of a default is equivalent to the simplified context (i.e., statements with variables connected to the random variables of the default). In the above, the simplified context is the conjunction of terms in

$$\beta \wedge \delta \wedge \forall \vec{\mathbf{z}}((\alpha \wedge \beta) \rightarrow \gamma) \wedge [\alpha|\beta]_{\vec{\mathbf{x}}} > \mathsf{c}$$

having free variables from $\vec{x}$. If the lemma applies then δ has no $x_i \in \vec{x}$ and since all other terms except β are closed formulas, the simplified context will be β. Since β is equivalent to the context of the default, the default applies.

To see how this works, consider Example 6.2. For the statistical knowledge base SKB6.2, the query

MTKB6.2:

fact elephant(clyde) & zookeeper(tony) & zookeeper(fred).
default elz(X,Y): likes(X,Y) ← elephant(X) & zookeeper(Y).
default ¬elf(X): ¬likes(X,fred) ← elephant(X).

SKB6.2:

$\texttt{elephant(clyde)} \;\&\; \texttt{zookeeper(tony)} \;\&\; \texttt{zookeeper(fred)},$
$[\texttt{elz(X,Y)}|\texttt{elephant(X)} \;\&\; \texttt{zookeeper(Y)}]_{<X,Y>} > c,$
$\forall \texttt{X,Y.(elz(X,Y)} \;\&\; \texttt{elephant(X)} \;\&\; \texttt{zookeeper(Y))} \rightarrow \texttt{likes(X,Y)},$
$[\neg\texttt{elz(X)}|\texttt{elephant(X)}]_{X} > c,$
$\forall \texttt{X.(}\neg\texttt{elf(X)} \;\&\; \texttt{elephant(X))} \rightarrow \neg\texttt{likes(X,fred)}.$

Example 6.2: Elephants and Zookeepers

$$\mathrm{prob}\Big(\neg\texttt{likes(clyde,fred)}\Big)$$

is evaluated by determining

$$E\Big([\neg\texttt{likes(clyde,fred)}^{v}|\texttt{SKB6.2}^{v}]_{\vec{v}}\Big)$$
$$= E\Big([\neg\texttt{likes(U,V)}|\texttt{elephant(U)} \;\&\; \texttt{zookeeper(V)} \;\&$$
$$[\neg\texttt{elz(X)}|\texttt{elephant(X)}]_{X} > c \;\&$$
$$\forall \texttt{X.(}\neg\texttt{elf(X)} \;\&\; \texttt{elephant(X))} \rightarrow \neg\texttt{likes(X,V)}]_{\langle U,V\rangle}\Big)$$

and, by Lemma 39, this is greater than c.

In evaluating the query ¬likes(clyde,fred) in MTKB6.2, the context is

$$\texttt{elephant(U)} \;\&\; \texttt{zookeeper(V)},$$
$$\forall \texttt{X.(}\neg\texttt{elf(X)} \;\&\; \texttt{elephant(X))} \rightarrow \neg\texttt{likes(X,V)}$$

where U and V are the random variables. The simplified context (when considering the default instance ¬elf(clyde)) is just elephant(U) because terms not connected (as per Figure 18) to the random variable U corresponding to the random variable of the default are eliminated. Since the simplified context is equivalent to the randomized context of the default ¬elf(clyde), namely, elephant(U), the default applies and hence ¬likes(clyde,fred) is predicted.

For the query, likes(clyde,fred), the simplified context is

```
elephant(U) & zookeeper(V),
∀X.(¬elf(X) & elephant(X)) → ¬likes(X,V)
```

since both U and V are random w.r.t. the default elz(clyde,fred). But since this is not equivalent to the randomized context of the default, namely, elephant(U) & zookeeper(V), the query is (correctly) not predicted.

So far, we have consider only the case where theories involve only one assumption. Taking other assumptions into account is a fairly simple matter. Other assumptions are analogous to conditioning terms in probability expressions (recall the interpretation of **explain** in section 6.1). So to incorporate other assumptions, we simply include them as part of the context of the query. The resulting pruning criterion is:

```
meta fact prune([H|T]) ←
    default(H) &
    prediction(H,PH) &
```

```
context(H,CH) &
random(H,RVH) &
extract_context(C0) &
¬equivalent_context(RVH,PH,CH,C0).
```

This involves modifying the meta-predicate *extract_context* to include the other assumptions T. In the algorithm of Figure 18, C_0 becomes the randomized facts plus the randomized assumptions T. For example, suppose in Example 6.2 we are interested in the query `likes(dumbo,tony)` given that we have already assumed `elephant(dumbo)`—say, because we had a conjecture `elephant(X)`. The simplified context in this case is

```
elephant(U) & zookeeper(V)
```

since both `U` and `V` are random w.r.t. the default `elz(dumbo,tony)`. Since this is equivalent to the randomized context of the default, the default applies and hence `likes(dumbo,tony)` is predicted.

Note that for the query `¬likes(dumbo,tony)`, although the simplified context is `elephant(U)` which is equivalent to the randomized context of `¬elf(dumbo)`, the query is (correctly) not predicted because `¬elf(dumbo)` & `elephant(dumbo)` implies `¬likes(dumbo,fred)` and not `¬likes(dumbo,tony)`.

6.3 Specifying the Reference Class Selection Policy

In this section, an approximation of the reference class selection policy of the previous chapter is developed. This approximation is specified as pruning criteria for defaults. A default is deemed inapplicable to a particular case by the pruning criteria if inheritance between the context of the default and the context of the particular case is not sanctioned by the reference class selection policy. This results in a system that corresponds to an extension of T_0 from Chapter 4 for the restricted case where all known statistical probabilities are greater than some unspecified constant c, less than $1 - c$, or equal to one or zero. This extension involves incorporating the Second Order Direct Inference Principle (Definition 37) as described in the previous chapter.

Though, as mentioned in the previous chapter, it is unclear, in general, how to compute interferences in applying the Second Order Direct Inference Principle, the reference class selection policy can be approximated by identifying characteristic patterns of interference. The conjectured pattern described below has been shown to hold in several small examples by exhaustive enumeration and has been shown statistically significant in larger examples by random sampling.

Suppose we are interested in the inheritance of a statistic from one context to another. For instance in SKB5.4, we are interested in the inheritance of the proportion with immunity (75%) from the context of vaccinated people to the context of vaccinated children. First, let us define some notions of connectedness between contexts. Contexts, such as vaccinated children and vaccinated people people in SKB5.4, that stand in the subset-superset relation are potential participants in inheritance (i.e., they may stand in the average subset-superset relation) and so we say they are connected contexts. Formally,

Definition 40 (Connected Context)

The context C_1 is said to be a *connected context* to the context C_2 if $C_1 \subseteq C_2$ or $C_2 \subseteq C_1$.

If two contexts are connected, the influence of one on the other may be interfered with by a third mediating context, that is, a closer connected context screens the influence of a more distant context (as well, a connected context is considered closer than an unconnected one). Vaccinated children (in SKB5.4) is a closer connected context to vaccinated people than is blonde vaccinated children. Closeness is formally defined as follows.

Definition 41 (Closer Connected Context)

The context C_3 is said to be a *closer connected context* to the context C_2 than the context C_1 if $C_1 \subset C_3 \subseteq C_2$ or $C_2 \subseteq C_3 \subset C_1$ (or if C_1 is not connected to C_2 while C_3 is connected to C_2).

In determining inheritance, we are interested in the closest neighbours of a context, that is, the closest connected contexts with known statistics, because these screen the influence of other contexts. The context vaccinated children is screened by its closest neighbours vaccinated people and blonde vaccinated children in SKB5.4. This is formalized below.

Definition 42 (Neighbour Screen)

The *P-neighbour screen* of a context C_1 is the set of closest connected contexts for which the statistic for property P is known. That is, C_2 is in the P-neighbour screen of C_1 iff C_2 is a connected context of C_1, the statistic for P in C_2 is known, and there is no closer connected context C_3 for which the statistic for P is known.

In SKB5.4, since blonde vaccinated children is in the *immune*-neighbour screen of vaccinated children and the immunity statistic differs from that in the context of vaccinated people (also in the *immune*-neighbour screen of vaccinated children), inheritance of the immunity statistic from vaccinated people to vaccinated children is not viable because the conflicting influence of the immunity statistic for blonde vaccinated children is not screened from vaccinated children. Similarly, inheritance from blonde vaccinated children to vaccinated children is not viable. Viable inheritance is formally defined as follows.

Definition 43 (Viable Inheritance)

Inheritance of the statistic S_2 for a property P from a context C_2 to a context C_1 is *viable* iff

1. C_2 is in the P-neighbour screen of C_1 and
2. there is no context C_3 with statistic S_3 for property P in the P-neigbour screen of C_1 where $S_2 \neq S_3$.

Conjecture 44 (Viable Inheritance Correspondence Conjecture) Inheritance of the statistic for α from context γ to context β is viable only if

$$\mathsf{E}\Big([\alpha|\beta]_{\mathsf{V}}\Big) = \mathsf{E}\Big([\alpha|\gamma]_{\mathsf{V}}\Big)$$

according to the Second Order Direct Inference Principle.

The converse, however, is false since the expectation equality may hold even when γ is not in the α-neighbour screen of β. If the conjecture is correct then the following pruning criterion captures the Second Order Direct Inference Principle.

```
meta fact prune([H|T]) ←
    default(H) &
    context(H,CH) &
    prediction(H,PH) &
    random(H,RVH) &
    extract_context(C0) &
    ¬viable_inheritance(RVH,PH,CH,C0).
```

Note that this is the pruning criterion from the previous section with the check for equivalent contexts replaced by a check for viable inheritance between the contexts. The meta-predicate *viable_inheritance* embodies Definition 43—the parameter PH, the property being inherited, is included because viable inheritance is defined relative to a particular property PH and its PH-neighbour screen.

```
meta fact viable_inheritance(RVH,PH,CH,C0) ↔
    in_neighbour_screen(RVH,PH,CH,C0) &
    ¬conflicting_neighbour_screen(RVH,PH,C0).
```

If we let CHV and PHV be respectively CH and PH randomized w.r.t. RVH then the meta-predicate *in_neighbour_screen* checks whether CHV is in the PHV-neighbour screen of $C0$ (see the first condition for viable inheritance in Definition 43) and the meta-predicate *conflicting_neighbour_screen* checks whether the PHV-neighbour screen of $C0$ contains contexts that disagree on the statistic for PHV (see the second condition for viable inheritance in Definition 43).

Unfortunately the specification of these two meta-predicates is non-trivial. To make some headway, an approximation is provided below. The empirical results in the next section indicate the merits and shortcomings of the approximation.

The meta-predicate *in_neighbour_screen* is defined according to Definition 42.

```
meta fact in_neighbour_screen(RV1,P1,C1,C0) ←
    known_statistic(RV1,P1,C0,C1,P1) &
    randomize(RV1,C1,C1V,[],V1) &
    simplify_context(RV1,P1,C0,Cn1) &
    connected_context(C1V,C0) &
    ¬∃ RV2,C2,P2,C2V,V2,RV,Cn
    ( known_statistic(RV2,P1,C0,C2,P2) &
    randomize(RV2,C2,C2V,[],V2) &
    simplify_context(RV2,P2,C0,Cn2) &
    connected_context(C2V,C0) &
    intersect(RV1,RV2,RV) &
    simplify_context(RV,P1,C0,Cn) &
    closer_connected_context(C2V,Cn,C1V) ).
```

The chief problem with this definition is specifying *known_statistic* which checks whether, in the context C, there is a statistic for P or its negation. A compromise solution is to consider only two cases. First, consider a context as having known statistics based on the facts. Second, consider the context of defaults as having known statistics—this unfortunately overlooks statistics inherited to other contexts (see the Albertan Blonde Vaccinated Children problem in Example 6.16).

The meta-predicates *connected_context* and *closer_connected_context* are straightforward. The reference to *prove* in the braces indicates a tie into the theorem prover (Figures 12 and 13). A means to link predicates to Prolog predicates is provided in Meta-Theorist (the details are not important for the current discussion). Note that *not_prove* is implemented via negation as failure in Prolog.

```
meta fact connected_context(C1,C2) ←
    { prove C1 from C2 } or
    { prove C2 from C1 }.

meta fact closer_connected_context(C3,C2,C1) ←
    ¬connected_context(C1,C2) &
    connected_context(C3,C2).
meta fact closer_connected_context(C3,C2,C1) ←
    connected_context(C1,C2) &
    { prove C2 from C3 } &
    { prove C3 from C1 } &
    { not_prove C1 from C3 }.
meta fact closer_connected_context(C3,C2,C1) ←
    connected_context(C1,C2) &
    { prove C3 from C2 } &
    { prove C1 from C3 } &
    { not_prove C3 from C1 }.
```

The second of the two meta-predicates in the definition of *viable_inheritance* is *conflicting_neighbour_screen.* This checks whether there are contexts in the PV-neighbour screen of $C0$ with different statistics.

```
meta fact conflicting_neighbour_screen(RV,P,C0) ↔
    in_neighbour_screen(RV,P,C1,C0) &
    negate(P,NP) &
    in_neighbour_screen(RV,NP,C2,C0).
```

The pruning criterion outlined in this section provides a specification of viable inheritance (modulo the difficulties with *known_statistic*) which is conjectured to capture the Second Order Direct Inference Principle. The examples in the next section give some empirical evidence to support this conjecture.

6.4 Examples

To explore the properties of the specification of viable inheritance given in the previous section, the pruning criterion was implemented in Meta-Theorist.[1] The examples in this section discuss the results of running various queries against several different Meta-Theorist knowledge bases. Except where noted, the queries resulted in predictions that are correct with respect to viable inheritance. Whether these results are also correct with respect to the Second Order Direct Inference Principle remains an open question.

[1] For efficiency, the actual implementation used Prolog directly rather than meta-facts.

```
MTKB6.3:

    fact bird(tweety).
    fact penguin(opus).
    fact bird(X) ← penguin(X).
    fact ¬fly(X) ← penguin(X).
    default birdsfly(X): fly(X) ← bird(X).
```

Example 6.3: Penguins Don't Fly

Example 6.3 describes a situation in which most birds fly and penguins do not fly. In this example, *fly(tweety)* and *¬fly(opus)* are predicted. In the case of *tweety*, the *¬fly* statistic for *penguin* does not interfere with the *fly* statistic for *bird* as *penguin* is not in the *fly*-neighbour screen of the context for *tweety*, namely, *bird*. This is because the context *bird* of the default *birdsfly* is closer to the context *bird* for *tweety* than is the context *penguin* which has the known statistic *¬fly*. A similar analysis applies to the prediction of *¬fly* for *opus*.

```
MTKB6.4:

    fact bird(tweety).
    fact penguin(opus).
    fact bird(X) ← penguin(X).
    default ¬penguinsfly(X): ¬fly(X) ← penguin(X).
    default birdsfly(X): fly(X) ← bird(X).
```

Example 6.4: Most Penguins Don't Fly

Example 6.4 is similar to the previous one except that, instead of all penguins not flying, we have that most penguins do not fly. As in the previous example, *fly(tweety)* and *¬fly(opus)* are predicted. The analysis is also similar. The only difference is that the known statistic for the context *penguin* is based on the default *¬penguinsfly* rather than on the facts. This difference, however, has no effect on the results.

```
MTKB6.5:

    fact bird(tweety).
    fact bird(X) ← penguin(X).
    fact ¬fly(X) ← penguin(X).
    default birdsfly(X): fly(X) ← bird(X).
```

Example 6.5: Negative Path

Example 6.5 illustrates what is called negative path reasoning where *¬penguin(tweety)* is predicted. This follows from *bird(tweety)*, most birds fly, fliers are not penguins, and First Order Direct Inference.

```
MTKB6.6:

    fact bird(tweety).
    fact penguin(opus).
    fact bird(X) ← penguin(X).
    default ¬penguinsfly(X): ¬fly(X) ← penguin(X).
    default birdsfly(X): fly(X) ← bird(X).
    fact wings(X) ← fly(X).
```

Example 6.6: Winged Fliers

In Example 6.6, *wings(tweety)* is predicted while *wings(opus)* is not. This is correct because of the fact *wings(X) ← fly(X)* implies that *wings* is at least as probable as *fly*.

```
MTKB6.7:

    fact quaker(nixon) & republican(nixon).
    default ¬rp(X): ¬pacifist(X) ← republican(X).
    default qp(X): pacifist(X) ← quaker(X).
```

Example 6.7: Ambiguous

Example 6.7 is a case where there is a conflicting neighbour screen. The context *quaker* & *republican* for *nixon* has a *pacifist*-neighbour screen containing the contexts *republican* and *quaker*. These contexts have conflicting statistics about *pacifist*. Consequently, inheritance of the *pacifist* statistic from either context to the context for *nixon* is not viable. The result is that neither *pacifist* nor ¬*pacifist* is predicted.

```
MTKB6.8:

    fact vaccinated(tom).
    fact vaccinated(mary).
    fact ¬immune(mary) ← blonde(mary).
    default iv(X): immune(X) ← vaccinated(X).
```

Example 6.8: Blonde Vaccinated People

In Example 6.8, the context for *mary* and for *tom* differs. In the case of *tom*, the context is *vaccinated* and so *immune* is predicted based on the default *iv*. In the case of *mary*, the context is

> *vaccinated* & (¬*immune* ← *blonde*).

The *immune*-neighbour screen is conflicting since it contains both the context *vaccinated* and the context

> *vaccinated* & (¬*immune* ← *blonde*) & *blonde*.

Consequently, neither *immune* nor ¬*immune* is predicted for *mary*.

```
MTKB6.9:

    fact vaccinated_child(mary).
    fact vaccinated_child(wendy) & ¬blonde(wendy).
    fact vaccinated(X) ← vaccinated_child(X).
    default ¬ibvc(X): ¬immune(X) ← blonde(X)
        & vaccinated_child(X).
    default iv(X): immune(X) ← vaccinated(X).
```

Example 6.9: Blonde Vaccinated Children

In Example 6.9, the context *vaccinated_child* for *mary* is wedged between the conflicting contexts *vaccinated* and *blonde* & *vaccinated_child*. For *mary*, neither *immune* nor ¬*immune* is predicted. In the case of *wendy*, the context is *vaccinated_child* & ¬*blonde*. The *immune*-neighbour screen contains *vaccinated* but not *blonde* & *vaccinated_child*. Consequently, inheritance from *vaccinated* is viable for *wendy* with the result that *immune(wendy)* is predicted.

```
MTKB6.10:

    fact penn_dutch_speaker(hermann).
    fact american(X) ← penn(X).
    fact german_speaker(X) ← penn_dutch_speaker(X).
    default pdp(X): penn(X) ← penn_dutch_speaker(X).
    default ¬gsa(X): ¬american(X) ← german_speaker(X).
```

Example 6.10: Pennsylvania Dutch

In Example 6.10, we have that *hermann* is a speaker of Pennsylvania Dutch, Pennsylvanians are Americans, Pennsylvanian Dutch speakers are German speakers, most speakers of Pennsylvania Dutch are Pennsylvanians, and most German speakers are not American. The context *penn_dutch_speaker* of the default *pdp* is a closer connected context to the *penn_dutch_speaker* for *hermann* than is the context *german_speaker* of the default ¬*gsa*. Consequently, inheritance from the context *german_speaker* is not viable while inheritance from the context *penn_dutch_speaker* is viable. The result is that *american(hermann)* is predicted.

```
MTKB6.11:

    fact royal_elephant(clyde).
    fact elephant(clyde).
    fact elephant(X) ← royal_elephant(X).
    default eg(X): gray(X) ← elephant(X).
    default ¬reg(X): ¬gray(X) ← royal_elephant(X).
```

Example 6.11: Redundant Information

In Example 6.11, the redundant information *elephant(clyde)* does not affect the inheritance of *¬gray* from the context *royal_elephant*. This is because the context *elephant(clyde) & royal_elephant* is equivalent to the context *royal_elephant*.

```
MTKB6.12:

    fact royal_elephant(clyde).
    fact african_elephant(clyde).
    fact elephant(X) ← royal_elephant(X).
    fact elephant(X) ← african_elephant(X).
    default eg(X): gray(X) ← elephant(X).
    default ¬reg(X): ¬gray(X) ← royal_elephant(X).
```

Example 6.12: Off Path Preemption

Example 6.12 illustrates that viable inheritance exhibits the behaviour called *off path preemption*. Here *¬gray* is predicted based on the context *royal_elephant* of the default *¬reg*. The property *african_elephant* does not interfere with this.

```
MTKB6.13:

    fact quaker(nixon) & republican(nixon).
    default ¬rp(X): ¬pacifist(X) ← republican(X).
    default qp(X): pacifist(X) ← quaker(X).
    fact conservative(X) ← republican(X).
    fact anti_military(X) ← pacifist(X).
    default ¬cam(X): ¬anti_military(X) ← conservative(X).
```

Example 6.13: Cascaded Ambiguity Propagation

Example 6.13 illustrates the ambiguity due to the conflicting *pacifist*-neighbour screen is propagated to the *anti_military*-neighbour screen. The result is that neither *anti_military* nor *¬anti_military* is predicted. However, if the ambiguity over *pacifist* is resolved (say by asserting either *pacifist* or *¬pacifist*) then *anti_military* is no longer ambiguous.

```
MTKB6.14:

    fact citizen(fred) & gullible(fred).
    fact crook(dick) & elected(dick).
    default ¬lcc(X,Y): ¬likes(X,Y) ← citizen(X) & crook(Y).
    default lgcec(X): likes(X,Y) ← gullible(X) & citizen(X) &
        elected(Y) & crook(Y).
```

Example 6.14: Relations

So far, the examples have involved only properties of individuals. Example 6.14 involves the relation *likes*. For the query about *likes(fred,dick)*, the context is *citizen(X) & gullible(X) & crook(Y) & elected(Y)*. The context of the default *lgcec* is a closer connected context than the context of the default *¬lcc*. Consequently, the prediction is *likes(fred,dick)*.

Example 6.15 is similar to Example 6.2. See the discussion in section 6.2 and note the subtle twist with respect to the constant *fred* in the default *¬tlf*. Here the additional fact *tall(fred) & friendly(clyde)* does not interfere with the inheritance so, as in Example 6.2, *likes(clyde,tony)* and *¬likes(clyde,fred)* are predicted.

```
MTKB6.15:

    fact elephant(clyde) & zookeeper(fred) & zookeeper(tony).
    fact tall(fred) & friendly(clyde).
    default ¬elf(X): ¬likes(X,fred) ← elephant(X).
    default el [illegible] X,Y) ← elephant(X) & zookeeper(Y).
```

E[illegible] .15: More Relations

```
MTKB6.16:

    fact vaccinated_child(mary) & albertan(mary).
    fact vaccinated(X) ← vaccinated_child(X).
    default ¬ibvc(X): ¬immune(X) ← blonde(X)
        & vaccinated_child(X).
    default iv(X): immune(X) ← vaccinated(X).
```

Example 6.16: Albertan Blonde Vaccinated Children

The final example in this section, Example 6.16, illustrates a problem with the specification of viable inheritance. The system incorrectly predicts *immune(mary)*. The correct result is that *immune(mary)* is ambiguous based on inheritance from the context *vaccinated_child* which has a conflicting *immune*-neighbour screen (as seen in Example 6.9). The incorrect prediction results from the naive specification of *known_statistic* which doesn't consider the context *vaccinated_child* as having a known statistic. Consequently, *vaccinated* is admitted to the *immune*-neighbour screen in spite of *vaccinated_child* being a closer connected context. Refining the specification of *known_statistic* is left for future work.

6.5 Conclusion

This chapter has provided a specification for defaults in Meta-Theorist via theory pruning criteria. The pruning criteria were intuitively justified by appealing to probabilities and direct inference as discussed in the previous two chapters. Second Order Direct Inference was specified indirectly via viable inheritance. The examples discussed in this chapter indicate the properties of this specification and provide some empirical justification.

Chapter 7

Conclusion

7.1 What has Been Accomplished

This dissertation has dealt with the problem of the semantics of defaults in a hypothetical reasoning framework. It was motivated by a desire to deepen our understanding of the nature of default knowledge and to provide a clear representation of this knowledge.

Logic and probability are used as tools in the analysis of default knowledge and in the specification of statistically motivated defaults. This has facilitated the representation of domain knowledge in a principled way. The examples provided in Chapter 6 illustrate the power and the limitations of the proposed specification.

An important contribution is the introduction of Second Order Direct Inference. Just as First Order Direct Inference bases the properties of particular individual on the properties of the set of similar individuals, Second Order Direct Inference bases the independencies of particular predicates on the independencies of the set of similar predicates. Both forms of direct inference hinge on the idea of randomization—a particular member of a class is assumed to have the properties of the average member of that class.

First and Second Order Direct Inference determine a reference class selection policy. This policy is more precisely specified than that of Reichenbach and it differs in its treatment of inheritance. In particular, cases, such as Example 6.9, where the relevant context is wedged between two conflicting contexts result in ambiguous inheritance by this policy while Reichenbach's principle fails to consider the conflict and sanctions downward inheritance.

While it appears to be difficult, in general, to compute Second Order Direct Inferences, a second contribution of this research is the Viable Inheritance Correspondence Conjecture. This suggests the possibility of computing Second Order Direct Inference by identifying the neighbour screen of a context and checking for conflicts. This method of determining viable inheritance underlies the implementation in Meta-Theorist described in Chapter 6.

This implementation constitutes a third contribution as it demonstrates the feasibility of the proposed representation of default knowledge and it allows for empirical study of the proposed reference class selection policy.

Additional contributions are the design and implementation of the Meta-Theorist hypothetical reasoning framework together with the two theorems that lead to incremental computation of theory pruning and preference.

7.2 Future Research

There are numerous directions—from theoretical extensions to practical applications—in which to pursue future research. The relationship between neighbour screens and *d-graph separation* in Pearl et al.'s theory of graphoids [56, 55], and the relationship between randomization and *maximum entropy* [28] seem to be promising avenues of future research. As well, the relationship between randomization and axiomatic characterizations of independence, such as [55, 12, 7], should be investigated. But perhaps the most interesting future direction is the application of statistically motivated defaults to temporal reasoning.

There is an extensive body of literature on the frame problem in temporal reasoning, i.e., the problem of succinctly representing and reasoning about non-change [4, 70, 23]. Much interest has been focused on nonmonotonic reasoning approaches and since Hanks and McDermott [20] discovered that these approaches give rise to the multiple extension problem, many solutions have been proposed (e.g., [29, 39, 78, 14, 21, 40]).

Amazingly, all these solutions fail to solve the extremely simple Russian Roulette Problem (see Example 7.1). In this example, a person is playing Russian roulette with a single chamber gun. It is not known whether the

MTKB7.1:

fact alive(0) & shoot(0).
fact ¬alive(T+1) ← shoot(T) & loaded(T).
default frame_alive(T): alive(T+1) ← alive(T).
default frame_not_alive(T): ¬alive(T+1) ← ¬alive(T)
default frame_loaded(T): loaded(T+1) ← loaded(T).
default frame_not_loaded(T): ¬loaded(T+1) ← ¬loaded(T)

Example 7.1: Russian Roulette Problem

gun is loaded but if it is loaded, the person will be dead after the shoot action. The four defaults are intended to represent the common sense notion of persistence, i.e., a property tends to be invariant for a typical action and situation. For instance, people tend to stay alive but some actions are atypical in that they result in death.

Intuitively, since we don't have any information about whether the gun is loaded, we have no reason to favour the belief that the person will remain alive over the belief that the person will die. Yet, all of the current approaches to the multiple extension problem favour the conclusion that the person lives!

For instance, the chronological maximization of persistence approach [14] selects the counterintuitive theory

$$T_1 = \{\text{frame\_alive(0)},\text{frame\_not\_alive(0)}, \text{frame\_loaded(0)},\text{frame\_not\_loaded(0)}\}$$

which entails *alive(1)* and ¬*loaded(0)*. The other chronological preference based approaches produce analogous results ([29, 39, 78]).

This particular problem is an instance of one of the three of problems pointed out by Haugh [21], i.e., chronological preference approaches fail when

- there is incomplete knowledge of the initial state;
- there is disjunctive knowledge of action effects; or
- there is observational knowledge of a non-initial state.

Statistically motivated defaults appear to correctly deal with each of these. Example 7.1 falls under the first case. The context for a query about *alive(1)* is $\{alive(T)\ \&\ shoot(T)\}$. This has a conflicting *alive(T+1)*-neighbour screen which contains the conflicting contexts

$$\{alive(T)\} \text{ and } \{alive(T)\ \&\ shoot(T)\ \&\ loaded(T)\}.$$

Because of the conflicting neighbour screen, inheritance of *alive* is not viable.

A closer examination shows that Example 7.1 is also an example of the second case. The fact

fact ¬alive(T+1) ← shoot(T) & loaded(T).

can be rewritten as

fact (¬alive(T+1) ∨ ¬loaded(T)) ← shoot(T).

An example of the third case is Kautz's Vanishing Car problem [29]. In this example, Henry parks his car at time 0. At time 2, he notices it is no longer in the parking lot. We are interested in whether the car is parked at

```
MTKB7.2:

    fact parked(0).
    fact ¬parked(2).
    default frame_parked(T): parked(T+1) ← parked(T).
    default frame_not_parked(T): ¬parked(T+1) ← ¬parked(T).
```

Example 7.2: Kautz's Vanishing Car Problem

time 1. We can represent this as in Example 7.2. The chronological preference approaches predict *parked(1)* but there is no intuitive justification for this—the car could have been towed away or stolen any time after it was last observed parked. Under the statistically motivated defaults approach, the context for a query about *parked(1)* is $\{$*parked(T)* & ¬*parked(T+2)*$\}$. This has a conflicting *parked(T+1)*-neighbour screen which contains the conflicting contexts $\{$*parked(T)*$\}$ and $\{$¬*parked(T+2)*$\}$. Because of the conflicting neighbour screen, inheritance of *parked* is not viable.

Another issue which immediately arises in temporal reasoning is the problem of chaining defaults. For instance, suppose we had a situation similar to the preceeding one except that the car was not observed at time 2. This is represented in Example 7.3.

Here we would like to predict that the car remains parked at time 1 and time 2 and so on indefinitely into the future. The prediction that the car is parked at time 1 follows from the statistical interpretation of the defaults. The prediction about time 2, however, does not follow. This is because the

```
MTKB7.3:

     fact parked(0).
     default frame_parked(T): parked(T+1) ← parked(T).
     default frame_not_parked(T): ¬parked(T+1) ← ¬parked(T).
```

Example 7.3: Parked Car Problem

context for a query about *parked(2)* is $\{$*parked(T)*$\}$ which has a conflicting *parked(T+2)*-neighbour screen. The conflicting contexts in the neighbour screen are $\{$*parked(T) & parked(T-1)*$\}$ and $\{$*parked(T) & ¬parked(T+3)*$\}$. Because of the conflicting neighbour screen, inheritance of *parked* is not viable so the prediction *parked(2)* is not sanctioned.

This is not a problem. It merely indicates that *parked(2)* cannot be predicted at the same level of strength as *parked(1)*. But what we would like is to be able to predict *parked(2)* with a smaller degree of belief.

In Bacchus's terms, we would like to allow a weakened form of transitivity, i.e., from a statistical KB such as SKB7.4, we would like to infer $\mathsf{prob}\big(\mathsf{b(k)}\big) > c$ and $\mathsf{prob}\big(\mathsf{c(k)}\big) > c^2$ [3, p. 171ff.].

Weak transitivity requires the following expectation independence assumption.

$$\mathsf{E}\Big([\mathsf{c(V)}|\mathsf{b(V)}\ \&\ \mathsf{a(V)}]_{\mathsf{V}}\Big) = \mathsf{E}\Big([\mathsf{c(V)}|\mathsf{b(V)}]_{\mathsf{V}}\Big).$$

(See Bacchus for a complete analysis [3, p. 171ff.].) It is very promising that this independence assumption appears to follow from second order random-

SKB7.4:

$[b(X)|a(X)]_X > c.$
$[c(X)|b(X)]_X > c.$
$a(k)$

Example 7.4: Weak Transitivity

ization.

What this means in practical terms is that we can chain defaults in making predictions of various levels of strength. For instance, in Example 7.3, we can predict *parked(0)* at full strength based on the empty theory; we can predict *parked(1)* at the strength $> c$ based on the theory $\{$*frame_parked(0)*$\}$; we can predict *parked(2)* at the strength $> c^2$ based on the theory $\{$*frame_parked(0)*, *frame_parked(1)*$\}$; etc.

Finally, in applying statistically motivated defaults to Hanks's and McDermott's Yale Shooting Problem [20], an important observation is the reappearance of the multiple extension problem. Example 7.5 is a simplified version of the Yale Shooting Problem that retains its essential features.

As expected, the default instance *frame_alive(1)* is inapplicable because the *alive(T+2)*-neighbour screen of the context

$\{$*loaded(T)* & *alive(T+1)* & *shoot(T+1)*$\}$

contains the conflicting contexts

$\{$*loaded(T)* & *alive(T+1)* & *shoot(T+1)* & *loaded(T+1)*$\}$

MTKB7.5:

fact loaded(0) & alive(1) & shoot(1).
fact ¬alive(T+1) ← shoot(T) & loaded(T).
default frame_alive(T): alive(T+1) ← alive(T).
default frame_not_alive(T): ¬alive(T+1) ← ¬alive(T).
default frame_loaded(T): loaded(T+1) ← loaded(T).
default frame_not_loaded(T): ¬loaded(T+1) ← ¬loaded(T).

Example 7.5: Simplified Yale Shooting Problem

and

{*alive(T+1)*}.

This means that *alive(2)* is not predicted.

But the default instance *frame_loaded(0)* is also inapplicable because the *loaded(T+1)*-neighbour screen of the context

{*loaded(T) & alive(T+1) & shoot(T+1)*}

contains the conflicting contexts

{*loaded(T) & alive(T+1) & shoot(T+1) & alive(T+2)*}

and

{*loaded(T)*}.

Consequently, *loaded(1)* is not predicted (which in turn means ¬*alive(2)* is not predicted).

The first reaction to this is that there must be a problem with statistically motivated defaults. But what really seems to be the problem is that some independence knowledge used to arrive at our intuitive conclusion $\neg alive(2)$ has not been represented. The independence knowledge is essentially this: the future is independent of the past given a sufficiently detailed present that is, Markov's Principle [9]. This principle is in fact built into the notion of situation in situation calculus. It also seems to be what justifies the chronological preference approaches [29, 39, 78, 14] in the cases for which they work and it seems to explain the cases which do not work, e.g., insufficiently detailed present in the Russian Roulette Example. It seems that Pearl had Markov's Principle in mind when he wrote that "interactions mediated via unconfirmed future events can be discounted" [54, p. 103].

In the Simplified Yale Shooting Problem, it seems that Markov's Principle can be used to eliminate the conflicting context

$$\{loaded(T) \;\&\; alive(T+1) \;\&\; shoot(T+1) \;\&\; alive(T+2)\}$$

from the $loaded(T+1)$-neighbour screen since $loaded(T+1)$ depends only on $loaded(T)$ by Markov's Principle. This would result in the intuitive prediction $loaded(T+1)$ and $\neg alive(2)$.

There are undoubtedly many other directions that can be pursued in the theory and application of statistically motivated defaults.

Bibliography

[1] R. Aleliunas. Mathematical models of reasoning: Competence models of reasoning about propositions in English & their relationship to the concepts of probability. Research Report CS-87-31, Department of Computer Science, University of Waterloo, Waterloo, Ontario, July 1987.

[2] F. Bacchus. Representing and reasoning with probabilistic knowledge. Research Report CS-88-31, Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, July 1988.

[3] F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge.* MIT Press, Cambridge, Massachusetts, 1990.

[4] F.M. Brown, editor. *Proceedings of the 1987 Workshop: The Frame Problem in Artificial Intelligence*, Los Altos, California, April 1987. Morgan Kaufmann.

[5] R. Carnap. *Logical Foundations of Probability Theory.* University of Chicago Press, Chicago, Illinois, 1950.

[6] P. Cheeseman. An inquiry into computer understanding. *Computational Intelligence*. 4(1):58–66. 1988.

[7] J.P. Delgrande. An approach to default reasoning based on first-order conditional logic. *Artificial Intelligence*. 36(1):63–90. 1988.

[8] Workshop on defeasible reasoning with specificity and multiple inheritance, April 1989. [St. Louis. no published proceedings].

[9] E.B. Dynkin. *Markov Processes - I*. Academic Press. New York. 1965.

[10] Ronald Fagin and Joseph Y. Halpern. Belief. awareness, and limited reasoning. *Artificial Intelligence*. 34(1):39–76. 1988.

[11] G. Ferguson. Identity and skolem functions in resolution-based hypothetical reasoning. Master's thesis. Department of Computing Science. University of Alberta. University of Alberta. August 1989.

[12] H. Geffner. On the logic of defaults. In *Proceedings of the Seventh National Conference on Artificial Intelligence*. pages 449–454. 1988.

[13] R. Goebel. A sketch of analogy as reasoning with equality hypotheses. In K. Jantke, editor. *Analogical and Inductive Inference*. volume 397 of *Lecture Notes in Computer Science*, pages 243–253. Springer Verlag, Berlin, 1989.

[14] R.G. Goebel and S.D. Goodwin. Applying theory formation to the planning problem. In *Proceedings of the 1987 Workshop: The Frame Problem in Artificial Intelligence*, pages 207–232, Los Altos, California, April 1987. Morgan Kaufmann.

[15] S.D. Goodwin. Representing frame axioms as defaults. Master's thesis, Department of Computer Science, University of Waterloo, Waterloo, Ontario, May 1987. [also Research Report CS-87-48].

[16] S.D. Goodwin and J.D.D. Gagné. Explanation and prediction. [unpublished], May 1987.

[17] S.D. Goodwin and R.G. Goebel. Theory preference based on persistence. Research Report CS-86-34, Department of Computer Science, University of Waterloo, Waterloo, Ontario, September 1986.

[18] S.D. Goodwin and R.G. Goebel. Nonmonotonic reasoning in temporal domains: The knowledge independence problem. In *Proceedings of the Second International Workshop on Nonmonotonic Reasoning*, pages 187–201, January 1989.

[19] J. Halpern. An analysis of first-order logics of probability. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1375–1381, August 1989.

[20] S. Hanks and D.V. McDermott. Default reasoning, nonmonotonic logics, and the frame problem. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 328–333, Los Altos, California, August 1986. Morgan Kaufmann.

[21] B. Haugh. Simple causal minimizations for temporal persistence and projection. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 218–223, Los Altos, California, July 1987. Morgan Kaufmann.

[22] P.J. Hayes. In defence of logic. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pages 559–565, Cambridge, Massachusetts, August 22-25 1977. MIT.

[23] P.J. Hayes and K. Ford, editors. *Advances in Human and Machine Coginition: The Frame Problem in Artificial Intelligence*. JAI Press, 1991. [in preparation].

[24] C. Hempel and P. Oppenheim. Studies in the logic of explanation. *Philosophy of Science*, 15, 1948.

[25] C.G. Hempel. *Aspects of Scientific Explanation and other essays in the Philosophy of Science*. The Free Press, New York, 1965.

[26] D.J. Israel. What's wrong with non-monotonic logic? In *Proceedings of the First National Conference on Artificial Intelligence*, pages 99–101, Stanford, California, August 18-21 1980. Stanford University.

[27] W.K. Jackson. A theory formation framework for learning by analogy. Master's thesis, Department of Computer Science, University of Waterloo, Waterloo, Ontario, December 1986.

[28] E.T. Jaynes. Where do we stand on maximal entropy? In R.D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*. MIT Press, Cambridge, 1979.

[29] H. Kautz. The logic of persistence. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 401–405, Los Altos, California, August 1986. Morgan Kaufmann.

[30] B. Kirby. Preferring the most specific extension. In *Experiments in the Theorist Paradigm: A Collection of Student Papers on the Theorist Project*. Department of Computer Science, University of Waterloo, May 1987. [Research Report CS-87-30].

[31] A. Kolmogoroff. *Foundations of the Theory of Probability*. Chelsea Publishing Company, New York, 1950.

[32] K. Konolige and K. Meyers. Representing defaults with epistemic concepts. *Computational Intelligence*, 5(1):32–44, 1989.

[33] H.E. Kyburg, Jr. *Probability Theory.* Prentice-Hall, Englewood Cliffs, New Jersey, 1969.

[34] H.E. Kyburg, Jr. *The Logical Foundations of Statistical Inference.* D. Reidel, Dordrecht, Netherlands, 1974.

[35] H.E. Kyburg, Jr. The reference class. *Philosophy of Science*, 50(3):374–397, September 1983.

[36] H.E. Kyburg, Jr. Probablistic inference and nonmonotonic inference. In *Proceedings Fourth AAAI Workshop on Uncertainty in AI*, pages 229–236, 1988.

[37] H. Levesque (ed.). Taking issue forum: A critique of pure reason. *Computational Intelligence*, 3(3):147-227, 1987.

[38] I. Levi. *The Enterprise of Knowledge.* MIT Press, Cambridge, Massachusetts, 1980.

[39] V. Lifschitz. Pointwise circumscription: Preliminary report. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 406-410, Los Altos, California, August 1986. Morgan Kaufmann.

[40] V. Lifschitz. Formal theories of action. In *Proceedings of the 1987 Workshop: The Frame Problem in Artificial Intelligence*, pages 35-57, Los Altos, California, April 1987. Morgan Kaufmann.

[41] D. Lin and R. Goebel. Computing circumscription of ground theories with Theorist. Technical Report TR 89-26, Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada, October 1989.

[42] D.W. Loveland. *Automated Theorem Proving: A Logical Basis*. North-Holland, Amsterdam, 1978.

[43] B. Magee. *Popper*. Fontana Press, London, England, 1973.

[44] J. McCarthy. Epistemological problems of artificial intelligence. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pages 1038–1044, Cambridge, Massachusetts, August 22-25 1977. MIT.

[45] J. McCarthy. Circumscription—A form of non-monotonic reasoning. *Artificial Intelligence*, 13(1 & 2):27–39, 1980.

[46] J. McCarthy. Applications of circumscription to formalizing commonsense knowledge. *Artificial Intelligence*, 28(1):89–116, 1986.

[47] J. McCarthy and P.J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969.

[48] D.V. McDermott. A critique of pure reason. *Computational Intelligence*, 3(3):151–160, 1987.

[49] D.V. McDermott and J. Doyle. Non-monotonic logic I. *Artificial Intelligence*, 13(1 & 2):41-72, 1980.

[50] M. McLeish (ed.). Taking issue forum: An inquiry into computer understanding. *Computational Intelligence*, 4(1):57–142, 1988.

[51] L. Naish. Negation and quantifiers in nu-prolog. In *Proceedings of the Third International Conference on Logic Programming*, pages 624–634, London, England, July 14-18 1986. Imperial College.

[52] E. Neufeld. Defaults and probabilities: extensions and coherence. In *Proceedings First International Conference On Principles of Knowledge Representation and Reasoning*, pages 312–323, Toronto, Ontario, May 1989.

[53] N.J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28:71–87, 1986.

[54] J. Pearl. On logic and probability. *Computational Intelligence*, 4(1):99–103, 1988.

[55] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, California, 1988.

[56] J. Pearl and T.S. Verma. The logic of representing dependencies by directed graphs. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 374–379, 1987.

[57] J. Pollock. A theory of direct inference. *Theory and Decision*, 15:29–96, 1983.

[58] J. Pollock. Foundations for direct inference. *Theory and Decision*, 17:221–256, 1984.

[59] D.L. Poole. On the comparison of theories: Preferring the most specific explanation. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 144–147, Los Angeles, California, August 16-18 1985. UCLA.

[60] D.L. Poole. Default reasoning and diagnosis as theory formation. Research Report CS-86-08, Department of Computer Science, University of Waterloo, Waterloo, Ontario, March 1986.

[61] D.L. Poole. Defaults and conjectures: Hypothetical reasoning for explanation and prediction. Research Report CS-87-54, Department of Computer Science, University of Waterloo, Waterloo, Ontario, October 1987.

[62] D.L. Poole. Fixed predicates in default reasoning. Research Report CS-87-11. Department of Computer Science, University of Waterloo, Waterloo, Ontario, February 1987.

[63] D.L. Poole. Obseravtion and prediction: Distinctions in hypothetical reasoning. [unpublished], September 1987.

[64] D.L. Poole. Compiling a default reasoning system into prolog. Research Report CS-88-01. Department of Computer Science, University of Waterloo, Waterloo, Ontario, January 1988.

[65] D.L. Poole. A logical framework for default reasoning. *Artificial Intelligence*, 36(1):27–47, August 1988.

[66] D.L. Poole. Explanation and prediction: An architecture for default and abductive reasoning. *Computational Intelligence*, 5(2):97–110, May 1989.

[67] D.L. Poole, R.G. Goebel, and R. Aleliunas. Theorist: A logical reasoning system for defaults and diagnosis. In N. Cercone and G. McCalla, editors, *The Knowledge Frontier: Essays in the Representation of Knowledge*, pages 331–352. Springer-Verlag, New York, 1987.

[68] D.L. Poole and S.D. Goodwin. A Theorist to Prolog compiler. [unpublished], August 1987.

[69] K. Popper. *The Logic of Scientific Discovery.* Harper & Row, New York, 1958.

[70] Z.W. Pylyshyn, editor. *The Robot's Dilemma.* Ablex Publishing, 1987.

[71] H. Reichenbach. *Theory of Probability.* University of California Press, Los Angeles, California, 1949.

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

[73] N. Rescher. *Scientific Explanation.* Collier-MacMillian Canada, Toronto, 1970.

[74] A. Sattar and R. Goebel. On the efficiency of logic-based diagnosis. In *Proceedings of the Third International Conference on Industrial and Engineering Applications of AI/Expert Systems,* volume 1, pages 23–31, Charleston, South Carolina, USA, July 15-18 1990. ACM Press.

[75] A. Sattar and R. Goebel. Using crucial literals to select better theories. *Computational Intelligence,* 7(1):11–22, February 1991. [Also University of Alberta Department of Computing Science TR-89-27].

[76] A. Sattar and R. Goebel. An incremental nonmonotonic theorem prover. In *Proceedings of the International Symposium on* **Computational In-**

telligence — *Heterogeneous Knowledge Representation Systems*, Milano, Italy, September 24-28, 1990.

[77] A. Sattar, S. Goodwin, and R. Goebel. What is in consistency tree: A uniform treatment of theory selection. In *Proceedings of Pacific Rim International Conference on Artificial Intelligence*, pages 126–131, Nagoya, Japan, Novemeber 1990.

[78] Y. Shoham. Chronological ignorance: Time, nonmonotonicity, necessity and causal theories. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 389-393, Los Altos, California, August 1986. Morgan Kaufmann.

[79] G. Smith. *Statistical Reasoning*. Allyn and Bacon, Needham Heights, Massachusetts, 1988.

[80] D.S. Touretzky, J. Horty, and R. Thomason. A clash of intuitions: The current state of nonmonotonic inheritance systems. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 476–482, 1987.

[81] M.H. van Emden and R. Goebel. Research at Waterloo on logic-mediated, knowledge-based, personal information systems. *Canadian Artificial Intelligence*, 8:26–29, 1986.

[82] B. Wong. Analogical reasoning based on hypothetical reasoning. Master's thesis, Department of Computing Science, University of Alberta, Edmonton, Alberta, November 1989.