Computer Science Department Technical Report
Cognitive Systems Laboratory
University of California
Los Angeles, CA 90024-1596

DEFAULT REASONING: CAUSAL AND CONDITIONAL THEORIES

Hector Geffner

December 1989
CSD-890065

Default Reasoning: Causal and Conditional Theories

Hector Geffner
November, 1989

Technical Report 137
Cognitive Systems Laboratory
Department of Computer Science
UCLA
Los Angeles, CA 90024

This report reproduces a dissertation submitted to UCLA in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Computer Science. This work was supported in part by National Science Foundation Grants IRI-8610155, IRI-8821444, IRI-15522.

Abstract

Reasoning with Defaults:
Causal and Conditional Theories

Hector Geffner
UCLA
1987

Defaults play a central role in commonsense reasoning, permitting the generation of useful predictions in the absence of complete information. These predictions are *non-monotonic* in the sense that they often have to be revised in light of new information. Attempts to represent and reason with defaults in AI, however, have encountered the problem of spurious arguments: arguments which rely on acceptable defaults but which support unacceptable conclusions.

In this work we develop a semantic account of default reasoning which addresses this problem. First, we interpret defaults as *conditional assertions,* and appeal to probability theory and preference-model logics to uncover the basic set of inferences that such reading implies. Next, we extend the language of default theories to accommodate a causal operator which is used to distinguish *explained* from *unexplained* "abnormalities." Competing scenarios that arise from conflicting defaults are then rated in terms of the type of abnormalities they introduce, and the most *coherent* scenarios are preferred. The resulting framework yields a reasonable behavior in several domains of interest to AI, including inheritance hierarchies, reasoning about change, general logic programs and abductive reasoning.

Contents

List of Figures

Acknowledgements

Jacobo Sclarsky, my uncle Lito, started all this. He got me interested in the world of science before I ever went to school and has been my best teacher ever since.

For my years at UCLA, I want to thank first and foremost Judea Pearl, my thesis advisor. Judea taught me everything I know about research and provided me with that unusual combination of freedom and financial support without which this work would not have been possible. Judea is also responsible for making probability theory into a powerful conceptual framework from which the main ideas of this dissertation are drawn. He also provided valuable comments on an earlier draft and suggested significant improvements.

I am also grateful to the other members of my committee: Michael Dyer, Kit Fine, Keith Holyoak and Stott Parker for their support and suggestions. Kit Fine was especially generous with his time, providing sound criticism, insightful comments and a healthy dose of skepticism.

I also want to thank Gina George, Verra Morgan, Rosemarie Murphy, Doris Sublette and Judy Williams for bearing with me during all these years. I am also grateful to Tom Verma for discussions and good company, and to Bill Dolan for help with the presentation.

Many people outside UCLA provided comments, encouragement or both. Special thanks go to David Etherington, Matt Ginsberg, Ben Grosof, Ronald Loui and David Poole. I am particularly grateful to Ron for being the first person outside my family to like a piece of research of mine.

If I have kept my sanity after all these years it is only because of my wife, Maria Eugenia Fuenmayor. She has been a constant source of support and inspiration. To her, and to my family in Argentina who kept asking me "How come you haven't finished yet?" my deepest thanks of all.

La tesis esta dedicada a la memoria de mi madre, Sara Sclarsky, y de mis amigos Judith Goldberg, Mario Geffner y Ruben Gerenschtein.

– And where is the bellybutton?
– He doesn't have a bellybutton, Gille; because he hatched from an egg.

– And then his wings?
– He doesn't have wings either.
– How come. Didn't he hatch from an egg?

Yes, right, but not everything that hatches from an egg has wings. Lots of things come from eggs, like fish and spiders, and snakes, and birds, and ants, and frogs, and who knows what else.

Gee, eggs are really mixed up!

Joaquin Lavado (Quino)
Mafalda 7
Ediciones La Flor

Chapter 1

Introduction

1.1 Overview

The comic strip on the opposite page illustrates two pervasive aspects of commonsense inference: the elaboration of predictions in the absence of complete information and the ability to revise and explain predictions found to be wrong. Both aspects are so entrenched in common discourse that normally we forget that most of our actions are adopted on the basis of partial information and tentative beliefs. We get up in the morning and expect to find the coffee machine in the same place, the newspaper under the door, the tooth-paste in its container. But the coffee machine is not always in the same place, the newspaper not always under the door, tooth-paste not always in the tube. Still, these predictions are usually true and permit us to make plans that work most of the time. When they are not true, we adopt new beliefs leading to alternative actions and the old predictions are discarded.

Ubiquitous as these forms of reasoning are, they have resisted a satisfactory explanation. Why are both expectations "animal hatched form eggs have wings" and "reptiles have no wings" right, even though reptiles do hatch from eggs? Clearly, there is a high proportion of winged animals among those that hatch from eggs, yet a low proportion among reptiles. Still, the explanation of such expectations in terms of probabilities is not completely satisfying. These expectations rather appear to rely on *qualitative rules,* like "animals hatched from eggs have wings" and "reptiles do not have wings." Such rules, called *default rules,* express what is normally the case without ruling out the possibility of exceptions: turtles which

are hatched from eggs but do not have wings, pterodactyls which are reptiles but do have wings, and so on.

In Artificial Intelligence (AI), it has been natural to express commonsense knowledge in terms of defaults. Inheritance hierarchies, for instance, encode the prototypical properties of classes by means of defaults. In reasoning about change, defaults encode the tendency of properties to remain invariant in the absence of relevant changes. In diagnostic reasoning, defaults encode the absence of pathological behavior, whose presence must then be explained by postulating appropriate hypotheses. Even deductive databases usually embed default assumptions to fill in information not in the database.

However, attempts to represent and reason with defaults have encountered the problem of spurious arguments: arguments which rely on acceptable defaults but which support unacceptable conclusions. For instance, the argument that penguins fly, on the grounds that penguins are birds and birds normally fly, is not acceptable. Still, the same argument is acceptable about canaries. If knowledge representation languages are to accommodate default rules, then the *logic of default arguments* must be understood.

The language of logic appears as the most suitable candidate for describing the logic of default arguments. Precise and well-understood, a logical account of default inference should make explicit the criteria that distinguish good from bad default arguments, independently of domains and implementations. Indeed, logic itself was developed to describe sound argumentation. However, while logic is concerned with arguments that yield true conclusions from true premises, default reasoning is concerned with arguments that yield likely propositions from likely premises.

The first attempts in AI to provide a logical account of default inference focused on extending classical logic with *non-monotonicity*. Default reasoning is non-monotonic in the sense that default predictions often need to be revised in the light of new information. For example, we may go home early to watch the Lakers, only to discover, in the middle of a traffic jam, that we did not really leave early enough. Classical logic, on the other hand, is monotonic; no additional information can affect the status of a conclusion which is supported by a valid deductive argument.

The efforts to extend logic with non-monotonic features resulted in various non-monotonic formalisms which enabled certain patterns of default inference to be formulated in precise terms. These formalisms, for example, can support the

conclusions that somebody must be at home when the lights are on, switch to the opposite conclusion when nobody answers the bell, and switch once again to original conclusion when voices are heard through the window. This behavior is accomplished by regarding defaults as rules which extend a set of beliefs in the absence of conflicting evidence. However, while successful in accommodating patterns of default inference, non-monotonic logics do not uncover the logic of such patterns: a set of defaults normally gives rise to a number of conflicting default arguments, and non-monotonic logics leave it up to the user to distinguish the good from the bad.[1]

To account for the distinction between intuitive and counterintuitive arguments we not only need a non-monotonic logic but an interpretation of defaults capable of capturing and explaining default behavior. In this work we develop such an interpretation. For that we rely on two notions which have so far not been considered essential for understanding defaults. The first is the notion of *conditionals*. Conditionals are normally expressed in English by the form 'if A then B,' and are currently understood as context-dependent assertions.[2] They assert that B is true in a context defined by A. Conditionals with false antecedents, such as: "if I were not writing these lines I would be watching Crimes and Misdemeanors," are called *counterfactuals*. While counterfactuals are bound to be trivially true in classical logic, they may be false when analyzed conditionally. In such case, the truth of the counterfactual results from evaluating the truth of its consequent in a context in which its antecedent is true, and which preserves the relevant features from the current context.[3] Here we will adopt a *conditional interpretation of defaults:* a default 'if A then B' will be understood as asserting the truth of B in the context that results from the assimilation of A in a given *background context*. As we will show, such a view will have a definite impact on the type of default behavior which is legitimized.

The second thread in the proposed interpretation of defaults comes from the notions of *causality* and *explanations*. Defaults encode expectations, and violations of defaults represent expectation failures. The task of default reasoning is normally associated with the minimization of expectation failures. Such a view is most explicit in McCarthy's [1986] account of defaults, where default violations are encoded by means of "abnormality" predicates whose extensions are supposed to be minimal. Here we take a slightly different approach. Rather than treating

[1] Hereafter, the "user," refers to the builder of the knowledge base.

[2] See [Nute, 1984] for a survey on conditional logics.

[3] There are a number of important problems in determining what these relevant features are. See for instance, Goodman [1955].

all expectation failures in the same way, we distinguish between those which are *explained* from those which are *not explained.* The task of default inference is then associated with the minimization of *unexplained* expectation failures as opposed to the minimization of *all* expectation failures. In terms of McCarthy's abnormality formulation, this amounts to consider the "abnormality" of scenarios as opposed to the "abnormality" of individuals. So, for instance, no penalty will be associated with a scenario involving a *dead* non-flying bird called Tim, because even though Tim might be an abnormal bird, it is certainly not an abnormal *dead* bird, and that is what we claim matters.

Where do explanations come from? Usually they arise from *causal* relations. A moving action explains the change in location of a block, a disease explains a given symptom, and the termination of all vital activities explains a non-flying dead bird. Sometimes, however, the causal origin of explanations is not so apparent. Being a penguin explains being a non-flying bird, being a priest explains being an unmarried adult, and even Tom's desire to get home in time for a game explains him leaving earlier than usual. In each case, however, we will have a language capable of accommodating such explanatory patterns, which will permit us to assess the coherence of the competing scenarios that arise from conflicting defaults.

1.2 Representing Knowledge

A common current agreement in AI is that programs capable of reasoning about the world must embed large amounts of knowledge. Namely, knowledge about the world must be represented in the programs, and the behavior of such programs is to be explained in terms of the knowledge they embed. There are, however, many ways in which a program can embed a particular piece of knowledge. At one extreme knowledge can be embedded in the body of procedures, and at the other it can be encoded in declarative chunks with no commitment at all about its potential uses.

In AI it has been found useful to represent knowledge in *declarative* form. AI programs contain a set of expressions called the *knowledge base*, which are regarded as being in correspondence with the world represented, and a general purpose interpreter, often called the *inference engine*, which assembles the expressions in the knowledge base according to the goals at hand. This declarative organization of programs originated from the need to deal with ill-understood problems for which conventional top-down software techniques were not appropriate [Doyle, 1985], and

from the desire to make AI programs extensible [McCarthy, 1987].

Whether these features make the knowledge-based approach the most convenient paradigm for *constructing* intelligent programs — see the volumes by Rumelhart [1986], McClelland [1986] and others, for an emerging alternative view — the knowledge-based approach remains so far unchallenged as an approach to *understanding* intelligent programs and intelligent behavior in general.

A knowledge based program is determined by the *content* and the *interpretation* of its knowledge base, and two clearly different traditions have developed in AI for addressing them.

The "scruffy" approach, best represented by Schank's school, emphasizes the organization of the knowledge base for simulating how people process high-level information (see for instance, [Schank and Abelson, 1977]). Given a particular task, certain knowledge structures are postulated, and an interpreter is designed which handles these structures in an intuitively satisfying way. Programs in this tradition have illustrated both the psychological appeal and the computational importance of the organization of knowledge in memory (see Dyer's[1983] BORIS and Kolodner's [1984] CYRUS, among others).

"Neats", on the other hand, have argued that interpreters tailored to particular tasks are likely to lack the flexibility needed to endow programs with commonsense [McCarthy, 1968]. They say that the range of reasoning patterns should not be limited a priori by a 'knowledge engineer,' but should be implicit in the correspondence between the expressions in the knowledge base and the world being represented. Thus, work along the "neat" track has proceeded in the development of formal languages in which fragments of world knowledge can be encoded, and formal semantics in which the meaning of such encodings can be made precise. The interpreter then is to derive new expressions from old ones in ways compatible which such meanings. For its precision and clarity, classical logic has constituted the language of choice, often extended to accommodate temporal and epistemic notions, and non-monotonicity (see [Moore, 1985a, Levesque, 1987]; [McDermott, 1982, Allen, 1984], and [McCarthy, 1980, McDermott and Doyle, 1980, Reiter, 1980]).

A severe limitation of the "neat" approach is that these different logical formulations do not determine what is *useful* for the interpreter to do, only what is *valid*. What is valid, however, may often be useless, and sometimes what is useful may turn out to be invalid. Even determining validity may sometimes be out of the question [Levesque and Brachman, 1987]. Thus, it is reasonable to believe

that programs capable of displaying commonsense will require both semantic and architectural considerations to be taken into account.

In this work the focus is on the semantic aspects of default reasoning. The remarks above should thus warn the reader that even a satisfactory *semantic* account of defaults will still leave us short of a complete account of default *reasoning*. Hopefully, however, we will be closer.

Why bother with a semantic account of defaults? Due to their role in common discourse, every reasonably expressive knowledge representation language must accommodate defaults, and the problem arises as to what type of inferences they legitimize. A semantic account will characterize such inferences by associating a precise "meaning" to defaults; namely, definite constraints on the world that the defaults are intended to reflect.

When is a semantic account of defaults satisfactory? For our purposes, a semantic account of defaults will be satisfactory when, given a sufficiently rich description of the domain of interest in terms of logical assertions and defaults, it is able to distinguish intuitive from counterintuitive default arguments, making it intelligible why an argument belongs to one or the other category.

1.3 A Reader's Guide

Traditional non-monotonic logics regard defaults as rules for *extending* a set of beliefs in the absence of conflicting evidence. Different logics enforce such view in different forms: those which are defined proof-theoretically, by relying on *consistency* notions; those which are defined model-theoretically, by relying on *minimality* notions. In the reminder of this chapter we review such logics together with the systems (databases, truth maintenance systems, logic programs, etc.) from which they draw their main intuitions.

There is however more to default reasoning than non-monotonicity, and more to defaults than the extensional view. In chapter 2 we show that it is possible to build an alternative interpretation of defaults by regarding defaults of the form "normally, if p then q" as licenses to assume the conditional probability of q given p arbitrarily high, short of being one. Such an interpretation, called ϵ-semantics, leads to a qualitative set of inference rules called the *core*, having virtues and limitations that are practically orthogonal to those of traditional non-monotonic logics. In particular, as a result of the context-sensitivity nature of conditional probabil-

ities, ϵ-semantics can resolve arguments of different 'specificity' (e.g., "penguins don't fly in spite of being birds"), though it fails to account for arguments involving "irrelevance" assumptions (e.g., concluding "*red*-birds fly" from "birds fly"). The next step is the adoption of an additional rule, called the *irrelevance* rule, whose role is to derive sensible assumptions about conditional independence from the information available in the knowledge base. We show that the core augmented by the irrelevance rule combine the benefits of the probabilistic and the extensional views of defaults, and illustrate the resulting behavior on a number of examples.

Chapter 3 focuses on an alternative, non-probabilistic semantics of the core. Such a semantics is structured around the notion of preferential entailment advanced by Shoham [1988], and further developed by Kraus *et al.* [1988], Makinson [1989], and Lehmann and Magidor [1988]. The idea is to use defaults to determine a preference relation on models and to identify the valid predictions of a theory as those that hold in its preferred models. We also analyze the relationship between ϵ-entailment and preferential entailment, and show that under suitable conditions, the core is not only sound with respect to them, but also complete.

The goal of chapter 4 is the development of an extended conditional interpretation of defaults which validates both the core and the irrelevance rule. This is accomplished within the framework of preferential entailment, except that now defaults dictate the preference relations on models via admissible (default) assumption *priorities*. The resulting semantics, called *conditional entailment*, has many elements in common with McCarthy's [1986] prioritized circumscription, yet priorities do not need to be given by the user but are automatically extracted from the knowledge base. An alternative sound proof-theory for conditional entailment is also developed which, unlike the system defined in chapter 2, is also complete.

Conditional entailment integrates both the *extensional* view of defaults common to traditional non-monotonic logics, and the *conditional* view resulting from the probabilistic and the preferential interpretations. Still, examples can be constructed — the most notorious being the Yale "shooting problem" [Hanks and McDermott, 1986] — in which conditional entailment fails to account for the intuitive behavior.

In chapter 5, we refine conditional entailment by introducing a distinction between *explained* and *unexplained* abnormalities. This is done by extending the language of default theories with a *causal operator* 'C,' such that the expression $C\alpha$ holds when an abnormality α is explained. However, rather than considering whether an abnormality is explained in a *model*, we consider whether an abnormality is explained in a set of models called a *class*, which groups together models

committed to a common set of assumptions. Classes are then rated in terms of the type of abnormalities they introduce, and the most *coherent* classes are preferred. Next, we illustrate how a variety of domains of interest in AI, such as inheritance hierarchies, reasoning about change, general logic programs, and causal networks accept a natural formulation in the resulting framework.

Finally, in chapter 6, we summarize the main contributions and discuss some open problems.

1.4 Non-Monotonic Systems

In this section we will review some standard systems and tasks which involve non-monotonic forms of reasoning. Such systems, while not always based on clear logical foundations, are sufficiently simple and well-understood as to provide a flavor for the type of inferences that an adequate account of default inference must accommodate and explain. They show that even without an adequate formalization, we know a lot about how legitimate default inference should look like. We consider databases, inheritance hierarchies, general logic programs, truth-maintenance systems, and time map management systems.

Databases

Databases are systems designed for the efficient storage and retrieval of information about objects and their relationships. A departmental database, for example, may contain a relation `teaches` with two tuples (`martin`, `pascal`) and (`kay`, `lisp`). Relations and tuples are normally understood as encoding ground atoms; in this case, the atoms `teaches(martin, pascal)` and `teaches(kay, lisp)`. So the answer `martin` to a query "who teaches `pascal` or `c`," is understood from the fact that the atomic encoding of the database sanctions the sentence `teaches(martin, pascal)`$\vee$ `teaches(martin, c)` as a theorem.

However, a logical understanding of databases requires more than ground atoms. Given the database above, for instance, conclusions such as "`kay` does not `teach` `pascal`" and "only `martin` `teaches` `pascal`" will also be supported, even though they do not follow from the atomic encoding.[4] To account for such conclusions,

[4]We will not be too concerned here with the specific manner in which such conclusions are actually supported by the database. In general, databases will not allow queries such as "who

the atomic encoding must be augmented with certain assumptions about names of objects and about the world the database is supposed to represent. These are the *unique names assumption*, by which individuals with distinct names are assumed distinct, the *domain closure assumption*, by which all individuals are assumed named, and the *closed world assumption*, by which it is assumed that there are no more instances of a relation than those deducible form the database [Reiter, 1984].

In the case of the database above, the unique names, domain closure and closed world assumptions amount to augmenting the encoding of the database with formulas such as:[5]

$$\texttt{martin} \neq \texttt{kay}\,,\ \texttt{pascal} \neq \texttt{lisp}\,,\ \texttt{martin} \neq \texttt{pascal}\,,\ \ldots$$
$$\forall x.\ x = \texttt{martin} \vee x = \texttt{kay} \vee x = \texttt{pascal} \vee x = \texttt{lisp}$$
$$\forall x, y.\ \texttt{teaches}(x, y) \Rightarrow (x = \texttt{martin} \wedge y = \texttt{pascal}) \vee (x = \texttt{kay} \wedge y = \texttt{lisp})$$

Provided with these assumptions, the conclusions supported by the database will now be theorems of the logical encoding. However, as assumptions, these formulas may turn out to be false. For instance, a second `pascal` class taught by `kay`, may be opened, rendering the above closed world assumption false. In such a case, the database will no longer support the conclusion "`kay` does not `teach pascal`" but, rather, its negation.

Note that the logical interpretation of the database is not incremental: the addition of new information not only translates into the addition of new formulas, but also in the replacement of old formulas by new ones. Such outcome should not be surprising though; the behavior of the database changes *non-monotonically*, while the behavior of its logical encoding can only change *monotonically*. It suggests, nonetheless, how classical logic could be extended with non-monotonic features: by means of a closed world assumption capable of adapting itself dynamically to the contents of the database. As we will see in section 1.5 below, this is indeed the main intuition behind circumscription.

Inheritance Hierarchies

Databases are designed with efficiency as a main concern. They usually store large amounts of data in a few fixed formats that permits fast storage and retrieval.

does not tech pascal." Such a query would normally have to be rephrased in a 'safe' form, such as "who, among the teachers, does not teach pascal" [Ullman, 1982].

[5]The symbol '$\Rightarrow$' is used to denote material implication.

Semantic networks, on the other hand, are focused both on computational and representational issues, providing more expressive languages in which knowledge can be encoded (see for instance, [Fahlman, 1979, Sowa, 1984, Brachman and Schmolze, 1985]). The central idea, which has enjoyed significant cognitive appeal, is to represent knowledge in terms of directed graphs, with links representing relations among concepts. Here we will be concerned with a restricted version of semantic networks, commonly referred to as inheritance networks, in which the only relation of interest is that of class inclusion (e.g. [Touretzky, 1986]).

Figure 1.1, depicts a simple inheritance network. The network involves two types of links: positive links ($\rightarrow$), which assert that one class is a (not necessarily strict) subclass of another (e.g. birds are flying things), and negative links ($\not\rightarrow$) which assert that one class is a (not necessarily strict) subclass of the *complement* of another (e.g. penguins are not flying things).

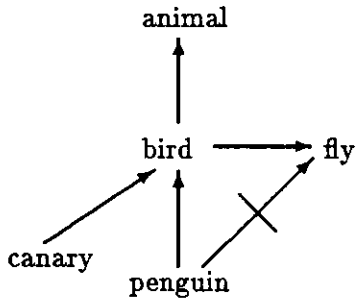


Figure 1.1: Simple Inheritance Hierarchy

Classes are assumed to *inherit* their properties from superclasses. unless otherwise specified. In the net depicted in fig. 1.1, for instance, canaries are assumed to inherit the property 'fly' from birds, just as penguins are assumed to inherit the property 'animal.' On the other hand, penguins do not inherit the property 'fly' from birds, as the link from penguins to the *negation* of 'fly,' being more "specific" than the link from birds to 'fly', overrides the inheritance path 'penguin' $\rightarrow$ 'bird' $\rightarrow$ 'fly.'[6]

Inheritance reasoning is also non-monotonic: a bird will normally be assumed to fly, though a penguin, which is also a bird, will not; more information thus results

[6]While in this case the choice is clear, the problem of determining "specificity" conditions in general networks has been a subject of much debate. See for instance, Touretzky [1986], Horty *et al.* [1987], and Geffner and Verma [1989] among others.

in the retraction of conclusions. Compared to databases, inheritance hierarchies point to additional aspects of non-monotonic reasoning that an adequate account of default reasoning must explain; in this case, the preference of more "specific" defaults over less "specific" ones.

Logic Programs

Logic programs are collections of implicitly universally quantified rules of the form $A \leftarrow L_1, L_2, \ldots, L_n$, where A is an atom, called the head of the rule, and each L_i, $i = 1, \ldots, n$, $n \geq 0$, is positive or negative literal in the rule's body. Unlike common programming languages constructs, the rules in a logic program accept both a procedural and a declarative reading [Kowalski, 1979]. If we restrict ourselves to propositional programs without negative literals, a rule $A \leftarrow L_1, L_2, \ldots, L_n$ can be understood both as stating that the goal A will be *derivable* when each of the subgoals L_i, $i = 1, \ldots, n$, is *derivable*, and that A is *true* when the literals L_i, $i = 1, \ldots, n$ are *true*.

When some of the literals L_i are negative, however, things are not so simple and the declarative reading of logic programs is usually dropped. Such programs are commonly understood in procedural terms, with the proviso that negative literals $\neg A_i$ are assumed to be derivable when every derivation for the atom A_i fails (see [Rousell, 1975]). Such form of negation has turned out to be a particularly useful programming tool, and follows a tradition that goes back to Planner-like languages [Hewitt, 1972]. Coined *negation as failure*, it endows logic programs with a behavior that is *non-monotonic*. In a program containing a single rule $\mathtt{p} \leftarrow \neg\mathtt{q}$, for instance, negation as failure yields a derivation for the atom p, which no longer holds when the rule $\mathtt{q} \leftarrow$ is added.

While the straightforward declarative reading of logic programs does not legitimize the behavior of negation as failure, more adequate logical accounts have been recently developed. In chapter 5, we will consider some of these accounts, as we analyze the relation between logic programs and *causal* default theories.

Truth Maintenance Systems

Truth maintenance systems (TMSs) are systems which keep track of dependencies among propositions [Doyle, 1979]. A user expresses justifications among beliefs in a restricted propositional language and the TMS labels each proposition as believed

(IN) or not believed (OUT), according to whether the proposition is justified or not. As an example consider the following propositions:

W: "John is at work"
H: "John is at home"
D: "Today is a working today"
C: "John's car is in the garage".

A typical Doyle's TMS will then contain *justifications* of the form:[7]

J_1 : $W \leftarrow\!\!\!| D, \neg H$
J_2 : $H \leftarrow\!\!\!| C$
J_3 : $D \leftarrow\!\!\!|$

The first justification J_1 can be read as saying that it is justified to believe that John is a work if it is believed that today is a working day and it is not believed that John is at home, while J_3 says that it is justified to believe that today is a working day.

The TMS algorithm labels propositions as believed (IN) when they are justified.[8] Given the justifications above, for example, the labeling algorithm will have both D and W labeled 'IN,' as D is justified as a premise, while W is justified by D and the lack of belief in H.

Such beliefs, however, are subject to revision. Consider for instance that we find John's car in his home's garage:

J_4 : $C \leftarrow\!\!\!|$

The belief that John is at home (H) becomes now justified and, as a result, the justification for the belief that John is at work (W) no longer holds, so the TMS deletes W from the IN list, and adds H.

Though understood in procedural terms for a long time, some satisfactory accounts of the semantics of the TMS belief revision process have been recently advanced [Elkan, 1988, Reinfrank *et al.*, 1989]. They are based on autoepistemic logic, a formalism for non-monotonic reasoning which we will discuss in the next section. More interestingly, such accounts reveal that a Doyle's TMS is not very different from a propositional logic program, and that a TMS labeling turns out to

[7]The syntax we use is different from Doyle's.

[8]There are subtle but important issues about circular justifications which we ignore here. See [Doyle, 1979] for details.

be nothing else but a *stable model* of the logic program that results from replacing the justification construct '$\leftarrow\!\!\!|$' by the logic programming construct '$\leftarrow$.'[9]

Time Map Management Systems

Time map management systems (TMMs) [Dean and McDermott, 1987] are systems for efficiently reasoning about propositions whose status changes in time. In one of its simplest forms, given a set of propositions which hold at a given time, a TMM infers the propositions that will hold at some later time, after a sequence of events has taken place [Dean and Boddy, 1987]. All such propositions are assumed to persist in the absence of relevant changes, while sources of change are encoded in the form of causal rules stating the effects of events when certain conditions are satisfied. A causal rule may indicate, for instance, that after checking a book out of the library, the book is no longer at the library, but in possession of the borrower; a second rule may state that if the borrower has checked a book at a some time t and has not returned it for a certain period of time T, s/he will get a fine at time $t+T+\Delta$, and so on.

In the simple case described, the task of the TMM is straightforward (see also [Hanks and McDermott, 1985]). The algorithm starts at the time for which it has complete information, and moves along the time axis looking for causal rules which may be triggered. If so, all such rules are inspected and the status of the temporal database is updated accordingly.

Simple as it is, however, this projection task is very instructive of what a system of default reasoning about change must be able to do. We will come back to these issues in section 5.2, when we will analyze some of the general requirements to be met by adequate frameworks for reasoning about change.

1.5 Non-Monotonic Logics

All the systems reviewed in the previous section behave non-monotonically. Still, such a non-monotonic behavior is the result of well-crafted algorithms. Non-monotonic logics were developed to understand what these algorithms do in logical

[9] For the stable semantics of logic programs, see [Gelfond and Lifschitz, 1988]. The correspondence between TMS labelings and stable models is elaborated in [Elkan, 1988]. We will say more about stable models and their equivalent felicitous models [Fine, 1989], in section 5.3.

terms, making explicit the assumptions they embed as well as their virtues and limitations. Moreover, while all these systems rely on languages with restricted expressive power, non-monotonic logics were developed to add non-monotonic behavior to languages with the expressivity of first order logic. In this section we will review three such standard non-monotonic formalisms: Reiter's [1980] default logic, McCarthy's [1980, 1986] circumscription, and Moore's [1985b] autoepistemic logic.[10]

Default Logic

In Reiter's [1980] default logic, defaults are tentative rules of inference of the form:

$$\frac{\alpha(x) : \beta(x)}{\gamma(x)}$$

where $\alpha(x)$, $\beta(x)$ and $\gamma(x)$ are formulas with free variables among those of $x = \{x_1, x_2, \ldots\}$, called the precondition, the test condition and the consequent, respectively. For a tuple a of ground terms, such a default permits one to derive $\gamma(a)$ from $\alpha(x)$, *provided that* $\neg\beta(a)$ *is not derivable.*

For instance, a default

$$\frac{\texttt{bird(x)} : \texttt{flies(x)}}{\texttt{flies(x)}}$$

permits the conclusion `flies(Tim)` upon learning `bird(Tim)`. However, if the negation of `flies(Tim)` is observed, the default gets blocked and the former conclusion no longer holds.

The appeal to non-derivability in the body of defaults together with their use to extend the set of derivable sentences, often leads to conflicts among defaults. For instance, given the additional default:

$$\frac{\texttt{injured(x)} : \neg\texttt{flies(x)}}{\neg\texttt{flies(x)}}$$

[10]For more detailed surveys on non-monotonic reasoning, see [Ginsberg, 1987, chapter 1], and [Reiter, 1987a]. McDermott [1987] also reviews some of these issues within the broader context of logic in AI.

and that Tim is also injured, we obtain a situation in which the preconditions of two defaults are satisfied, but in which the application of one blocks the other and vice versa. Reiter deals with such conflicts by introducing the notion of *extensions* of a default theory. Here, for simplicity, we will only consider the extensions of *normal* default theories, where the test conditions and consequents of defaults coincide.

An *extension* of a normal default theory $T = (W, D)$, where W is a set of wffs and D is a set of defaults, is a minimal deductively closed set of formulas F, $W \subseteq F$, such that every default $\alpha : \beta / \gamma$ in D whose precondition α is in F is either blocked, i.e. $F \vdash \neg\beta$, or has its consequent γ in F.

In the example above, two different extensions can be constructed: one which results from the application of the first default, corresponding to the theorems derivable from $\{\texttt{bird(Tim)}, \texttt{injured(Tim)}, \texttt{flies(Tim)}, \}$, and a second which results from the application of the second default, corresponding to the theorems derivable from $\{\texttt{bird(Tim)}, \texttt{injured(Tim)}, \neg\texttt{flies(Tim)}\}$.

Reiter's default logic main merit lies in extending classical first order logic with non-monotonic features by means of a simple formal device. Such an extension is sufficient to account for some of the non-monotonic forms of inference that arise in databases, and, by careful encoding, other forms of reasoning as well (see for example [Etherington and Reiter, 1983], for the encoding of inheritance hierarchies). As a framework for representing and reasoning with defaults, however, default logic is too weak. The natural encoding of a body of knowledge in the form of a default theory often gives rise to unreasonable extensions, which must then be pruned by the user by properly tuning the defaults' test conditions [Reiter and Criscuolo, 1983]. In this regard, Reiter's logic is more a precise language for specifying non-monotonic behavior, than an interpretation for uncovering the meaning of databases containing defaults. It is such an interpretation, however, what we are looking for.

Circumscription

Circumscription is a formal device which added to a first order theory asserts that the objects that can be shown to satisfy certain predicate P are the only objects that do [McCarthy, 1980, McCarthy, 1986, Lifschitz, 1988a].[11] For instance, from a database only including the fact $\texttt{Q(a)}$, the circumscription of Q yields the formula $\forall \texttt{x}.\, \texttt{Q(x)} \Rightarrow \texttt{x} = \texttt{a}$ as a conclusion. Thus, if b is an object different from

[11] See also the text by Genesereth and Nilsson [1987].

a, the circumscription of Q will permits us to jump to the conclusion $\neg Q(b)$. If Q(b) is learned, however, the previous conclusion would no longer hold, and the new conclusions would correspond to those derivable from the formula $\forall x.\, Q(x) \Leftrightarrow x = a \vee x = b$. Circumscription thus behaves as a powerful 'adaptable' closed world assumption, capable of dealing with theories richer than those expressible in databases.

Formally, if we let A(P) stand for a first order sentence containing the predicate P, and let $A(\Phi)$ denote the sentence that results from replacing all the occurrences of P by a predicate Φ with the same arity as P, the circumscription Circ[A(P); P] of P in A(P) can be expressed as the second order schema [McCarthy, 80]:

$$A(P) \wedge A(\Phi) \wedge \forall x.\, [\Phi(x) \Rightarrow P(x)] \Rightarrow \forall x.\, (P(x) \Rightarrow \Phi(x)).$$

The schema can be understood as stating that among the predicates Φ that satisfy the constraints in $A(\Phi)$, P is the strongest.

In order to see how circumscription works, let us consider the sentence A(Q) : Q(a), and let us substitute in the predicate $\Phi(x)$ the expression $x = a$. Such substitution yields the closed first order formula:

$$Q(a) \wedge a = a \wedge \forall x.\, [x = a \Rightarrow Q(x)] \Rightarrow \forall x.\, Q(x) \Rightarrow x = a,$$

which can be simplified to:

$$Q(a) \wedge [\forall x.\, Q(x) \Rightarrow x = a]$$

from which a minimal definition of Q follows:

$$\forall x. Q(x) \Leftrightarrow x = a.$$

Circumscription can also be understood from a model-theoretic perspective. In classical logic, a sentence *s* is said to be entailed by a sentence A(P) if *s* holds in every model of A(P). Circumscription weakens this condition: a proposition s is entailed by Circ[A(P); P] if and only if *s* holds in every model of A(P) *minimal* in P [McCarthy, 1980, Lifschitz, 1985].[12] A model *M* is minimal in P when there is

[12]The 'if' part requires the universal closure of the circumscriptive schema.

no other model which assigns a strictly smaller extension to P and which preserves from M the same domain and the same interpretation of symbols other than P.

Notice the way by which circumscription achieves non-monotonic behavior: given a set of axioms, circumscription picks up a minimal interpretation for some predicate(s) *subject to the constraints imposed by the axioms*. As the base of axioms changes, so does the minimal interpretation circumscription picks, and thus, the inferential import of the circumscriptive schema.

For instance, given $A(\mathtt{Q}) : \mathtt{Q}(\mathtt{a})$, the circumscriptive schema reduces to the formula $\forall \mathtt{x}.\, \mathtt{Q}(\mathtt{x}) \Leftrightarrow \mathtt{x} = \mathtt{a}$. Similarly, for $A(\mathtt{Q}) = \mathtt{Q}(\mathtt{a}) \wedge \mathtt{Q}(\mathtt{b})$, the circumscriptive schema reduces to $\forall \mathtt{x}.\, \mathtt{Q}(\mathtt{x}) \Leftrightarrow \mathtt{x} = \mathtt{a} \vee \mathtt{x} = \mathtt{b}$. In either case, provided that $\mathtt{c}$ is different from $\mathtt{a}$ and $\mathtt{b}$, $\neg\mathtt{Q}(\mathtt{c})$ can be inferred. On the other hand, if $\mathtt{b}$ is different from $\mathtt{a}$, $\neg\mathtt{Q}(\mathtt{b})$ follows in the first case, but not in the second.

This way of 'jumping to conclusions' stands in contrast with the way Reiter's default logic achieves the same effect. In default logic, the situation above would be represented by means of a default:

$$\frac{: \neg\mathtt{Q}(\mathtt{x})}{\neg\mathtt{Q}(\mathtt{x})}$$

which, given $\mathtt{Q}(\mathtt{a})$, will permit us to jump to $\neg\mathtt{Q}(\mathtt{c})$ directly, independently of whether, say, $\mathtt{Q}(\mathtt{b})$ holds or not. On the other hand, while Reiter's default logic permits inferring $\neg\mathtt{Q}(t)$, for *each* term t, $t \neq \mathtt{a}$, it does not authorize concluding, as circumscription does, that $\mathtt{Q}$ does not hold for *all* individuals different than $\mathtt{a}$, i.e. $\forall \mathtt{x}.\, \mathtt{x} \neq \mathtt{a} \Rightarrow \neg\mathtt{Q}(\mathtt{x})$.

While circumscription adds non-monotonic features to first order logic, it does not uniquely specify how defeasible knowledge should be encoded. For that purpose McCarthy [1986] introduced a convention by which Reiter's normal defaults like

$$\frac{\mathtt{bird}(\mathtt{x}) : \mathtt{flies}(\mathtt{x})}{\mathtt{flies}(\mathtt{x})}$$

are encoded in the circumscriptive framework as object-level formulas

$$\forall \mathtt{x}.\mathtt{bird}(\mathtt{x}) \wedge \neg\mathtt{ab}_\mathtt{i}(\mathtt{x}) \Rightarrow \mathtt{flies}(\mathtt{x}),$$

read as "every *non-abnormal* bird with respect to flying flies." Once defaults are so expressed, the expected behavior follows from circumscribing the $\mathtt{ab}_\mathtt{i}$'s predicates,

or as McCarthy says, from "minimizing abnormality." However, before that can be done effectively, a more powerful form of circumscription is needed.

To illustrate this need, consider the default above, and a bird called `Tim`. Intuitively we would expect the circumscription of $\mathtt{ab_i}$ to yield $\neg\mathtt{ab_i}(\mathtt{Tim})$ and, therefore, $\mathtt{flies}(\mathtt{Tim})$. However, circumscription as presented so far, does not yield such a conclusion. To show that, consider a model M of the sentence above in which $\neg\mathtt{flies}(\mathtt{Tim})$ holds and `Tim` is the only abnormal individual. If M is not a minimal model in $\mathtt{ab_i}$, then there must be a model M' which assigns a smaller extension to the predicate $\mathtt{ab_i}$, and which preserves the interpretation of all other symbols. This however, amounts requiring that M' satisfy the sentence above together with the literals $\mathtt{bird}(\mathtt{Tim})$, $\neg\mathtt{flies}(\mathtt{Tim})$, $\mathtt{ab_i}(\mathtt{Tim})$ which is not possible. Thus, M is a minimal model, and therefore the soundness of circumscription guarantees that the sentence $\mathtt{flies}(\mathtt{Tim})$ will not be sanctioned.

What is needed in such cases is a form of circumscription in which certain predicates can be minimized *at the expense* of others. A more recent form of circumscription, proposed in [McCarthy, 1986], permits precisely that. The circumscription $\mathrm{Circ}[\mathrm{A}(\mathrm{P},\mathrm{Z});\mathrm{P},\mathbf{Z}]$ of the predicate P in the sentence $\mathrm{A}(\mathrm{P},\mathbf{Z})$, where $\mathbf{Z}$ stands for a tuple of predicates allowed to vary in the minimization of P, is defined by the second order formula:

$$\mathrm{A}(\mathrm{P},\mathbf{Z}) \wedge \forall\Phi,\Psi\, \mathrm{A}(\Phi,\Psi) \wedge \forall\mathbf{x}.\,[\Phi(\mathbf{x}) \Rightarrow \mathrm{P}(\mathbf{x})] \Rightarrow \forall\mathbf{x}.\,[\mathrm{P}(\mathbf{x}) \Rightarrow \Phi(\mathbf{x})].$$

This formula is stronger than the previous one, permitting not only substitutions in place of P, but also in place of the predicates in $\mathbf{Z}$.

The expected conclusion $\mathtt{flies}(\mathtt{Tim})$ in the example above follows now by circumscribing the predicate $\mathtt{ab_i}$, allowing the predicate `flies` to vary . To see that it suffices to substitute $\Phi(\mathbf{x})$ by $\mathbf{x} \neq \mathbf{x}$, and $\Psi(\mathbf{x})$ by $\mathbf{x} = \mathbf{x}$.

This extended form of circumscription also accepts an appealing model theoretic interpretation. The circumscriptive schema $\mathrm{Circ}[\mathrm{A}(\mathrm{P},\mathbf{Z});\mathrm{P},\mathbf{Z}]$ sanctions as theorems the sentences that hold in all models of the sentence $\mathrm{A}(\mathrm{P},\mathbf{Z})$ which are *minimal in* P *with respect to* $\mathbf{Z}$ [Lifschitz, 1985, Etherington, 1988]. A model M of $\mathrm{A}(\mathrm{P},\mathbf{Z})$ is minimal in P with respect to $\mathbf{Z}$, if there are no other models M' of $\mathrm{A}(\mathrm{P},\mathbf{Z})$ which assign a smaller extension to P, and which preserve from M the same domain and the same interpretation of symbols other than P and *those in* $\mathbf{Z}$.

While the discussion above focused on the circumscription of a single predicate symbol, the generalization to many predicate symbols, known as *parallel circum-*

scription, is straightforward. More interesting is the case of *prioritized circumscription*, in which the user is allowed to specify a priority ordering among the circumscribed predicates [McCarthy, 1986, Lifschitz, 1985, Lifschitz, 1988a]. For instance, the circumscription Circ[A; $P_1 > P_2 > \ldots > P_n$;Z] of predicates P_1,P_2, ..., P_n in decreasing order of priority, translates into the conjunction of $n-1$ circumscriptions of the form Circ[A; P_i;**Z** $\cup$ $\{P_{i+1}, \ldots, P_n\}$] together with Circ[A; P_n;**Z**]. Namely, predicates with higher priority are circumscribed at the expense of predicates with lower priority. Though it is not clear in general how priorities among predicates are to be selected, some general guidelines in specific domains have been advanced (see for instance, [Lifschitz, 1988b] in the domain of logic programs, and [Krishnaprasad *et al.*, 1989], in the domain of inheritance hierarchies).

Due to its power and mathematical tractability, circumscription has become the most extensively studied non-monotonic formalism. As a framework for reasoning with defaults, however, circumscription shares the same limitation of default logic: the distinction between good and bad default arguments is left to the user, who remains responsible for explicating the relevant preferences. Moreover, the treatment of equality and universals if often less appealing in circumscription than in default logic. For instance, circumscription will legitimize counterintuitive conclusions such as "all birds fly' given a default "birds fly," which are not certified by default logic. Similarly, if Tim is a bird and does not fly, circumscription, unlike default logic, will not jump to the conclusion that Tweety flies, unless Tim and Tweety are known to be different individuals.[13] Circumscription, however, offers a more expressive language than default logic, in which priorities play a major role. Such a role will be analyzed in detail in chapter 4.

Autoepistemic Logic

Autoepistemic logic is a non-monotonic extension of classical logic, originally proposed by Moore [1985b] as a reconstruction of McDermott's and Doyle's [1980] non-monotonic logic. Since then, autoepistemic logic has received growing attention, having been studied by Marek [1986], Konolige [1988], and Gelfond [1989] among others.

Autoepistemic logic deals with *autoepistemic theories:* propositional theories[14]

[13]Both problems could be solved if rather than minimizing the *extension* of circumscribed predicates P, we minimize the set of atomic truths $P(a)$, for all tuples a of ground terms in the language. See the discussion in section 4.3.

[14]See Konolige [1988] and Levesque [Levesque, 1987] for first order extensions.

augmented by a belief operator L, where sentences of the form $\mathrm{L}\alpha$ are read as "α is believed.". The *stable expansions* of an autoepistemic theory T are defined as the sets of formulas $S(T)$ which satisfy the equation

$$S(T) = \mathrm{Th}(T + \{\mathrm{L}p : p \in S(T)\} + \{\neg\mathrm{L}p : p \notin S(T)\})$$

where $\mathrm{Th}(X)$ stands for the set of tautological consequence of X. Stable expansions are supposed to reflect possible states of belief of an ideal rational agent, closed both under positive and negative introspection [Moore, 1985b].

A default such as 'if it is a bird, it flies' which in McCarthy's 'abnormality' formulation will be encoded as a sentence $\mathtt{bird} \wedge \neg\mathtt{ab_i} \Rightarrow \mathtt{flies}$ with a circumscribed predicate $\mathtt{ab_i}$, can be encoded in autoepistemic logic as a sentence $\mathtt{bird} \wedge \neg\mathrm{L}\mathtt{ab_i} \Rightarrow \mathtt{flies}$. Then, given `bird`, the only autoepistemic expansion will contain the autoepistemic sentence $\neg\mathrm{L}\mathtt{ab_i}$, and therefore, the target sentence `flies`.

An autoepistemic theory may also have none or many stable expansions. For instance, a theory such as $T = \{\neg\mathrm{L}p \Rightarrow p\}$ has no stable expansions, while a theory $T' = \{\neg\mathrm{L}p \Rightarrow q, \neg\mathrm{L}q \Rightarrow p\}$ has two.

In general, autoepistemic logic regards literals of the form $\neg\mathrm{L}\alpha$ as assumptions, and unless a proof for α can be constructed from other beliefs, those assumptions will appear in every expansion. Under certain circumstances, as in a theory $T = \{\mathrm{L}p \Rightarrow p\}$, literals of the form $\mathrm{L}\alpha$ will also act as assumptions, though whether they should act so has been debated [Konolige, 1988].

Autoepistemic logic has been successfully applied to characterize the semantics of general logic programs [Gelfond, 1987, Gelfond and Lifschitz, 1988] and truth maintenance systems [Elkan, 1988]. Both characterizations are natural and simple, requiring one only to replace logic negation by autoepistemic negation; namely, literals of the form $\neg p$ are replaced by literals of the form $\neg\mathrm{L}p$.

Other appealing features of autoepistemic logic follow from its autoepistemic character: no other non-monotonic logic can distinguish between belief on a proposition, from lack of belief on its negation.[15] Autoepistemic logic does so, and makes the lack of belief the preferred belief state.

On the negative side, the problems of autoepistemic logic as a framework for

[15]Except approaches such as Sandewal's [1988] and Ginsberg's [1988], based on partial models and multivalued logic, respectively.

default reasoning are of two types. On the technical side, exceptions often give rise to theories which lack stable expansions. For instance, the autoepistemic encoding $\texttt{bird} \wedge \neg \texttt{Lab}_\texttt{i} \Rightarrow \texttt{flies}$ of a default "if it is a bird, it flies" will lack stable expansions given the exception $\texttt{bird} \wedge \neg \texttt{flies}$. Some proposals for dealing with such difficulties have been recently advanced in [Gelfond and Przymusinska, 1989] and [Konolige and Myers, 1989].

On the conceptual side, though better suited than default logic and circumscription for certain default reasoning tasks (see for instance, [Gelfond, 1989] and section 5.4 below), autoepistemic logic still carries some of their shortcomings. In particular, none of these formalisms is able to account for the preference of a default "if p and r then $\neg q$" over a conflicting default "if p then $\neg q$," nor they can detect any inconsistency between two defaults "birds fly" and "birds not fly." Indeed, these formalisms give us no insight on the empirical basis that makes the first a good default, and the second a bad one. These aspects are left outside the logic for the user to care for.

To account for these aspects of defaults, the notion of defaults needs to be taken more seriously; not merely as rules for extending beliefs, but as declarative constraints over states of affairs. The nature and logic that governs those constraints will then provide us with a more faithful interpretation of default reasoning.

Chapter 2

A System of Defeasible Inference Based on Probabilities

2.1 Introduction

Belief commitment and belief revision are the two main characteristics of default reasoning. Beliefs are adopted in the absence of complete information, and often have to be revised when new information becomes available. The main tools for formalizing these notions, logic and probability, present serious limitations. Classical logic cannot accommodate belief revision; new information can only add new theorems, never remove old ones. Probability theory, on the other hand, while able to revise old beliefs in the light of new evidence, does not tell us much about belief commitment: propositions are believed only to a certain degree, never accepted as true for practical purposes.

Recently there has been a renewed effort to enhance both formalisms in order to overcome these limitations. Those working within the probabilistic framework have devised 'rules of acceptance' which work on top of a body of probabilistic knowledge to create a set of believed, though defeasible, propositions (see [Loui, 1987b] for a review). Those working within the logic camp have developed 'non-monotonic' extensions of classical logic in which old 'theorems' may defeated by new 'axioms' (see section 1.5 above).

In comparison, the probabilistic approach has enjoyed a significant advantage over the logical approach. Given a body of probabilistic knowledge there is in

general no question about what its consequences are. The issue is rather what constitutes an adequate acceptance rule. Non-monotonic logics, on the other hand, have lacked such a clear empirical content. Not only has it been difficult to account for the conclusions implicit in a body of defaults (e.g. [Hanks and McDermott, 1986]), but even to identify what these conclusions ought to be (see, for instance, [Touretzky *et al.*, 1987], "A clash of intuitions ...").

On the positive side, as noted by [Glymour and Thomason, 1984] and [Loui, 1987b], the logical approach has shown that a *qualitative* account of non-monotonic reasoning not requiring either 'acceptance rules' or the expense and precision of computing with numbers, might be possible, and has even suggested ways in which such an account may proceed.

The goal in this chapter is to show that it is possible to combine the best of both worlds. We develop a system of defeasible inference which operates very much like natural deduction systems in logic but which can be justified on probabilistic grounds. The resulting system is closely related to a logic of conditionals developed by Adams [1966], as we interpret defaults of the form $p \rightarrow q$ as *infinitesimal high conditional probability statements*. However, high probability turns out not to be enough for our purposes. As we show later, some simple patterns of default inference, such as default chaining, fail to be sanctioned from such an account. Thus we extend the probabilistic interpretation with a syntactic account of *irrelevance* used to draw independence assumptions. This notion of irrelevance endows the resulting account with a dialectical flavor common to approaches in which defeasible reasoning is viewed as emerging from the interaction of competing arguments (e.g., [Poole, 1985, Nute, 1986, Horty *et al.*, 1987, Loui, 1987a, Pollock, 1988]).

Two important benefits result from viewing defaults as high conditional probability statements augmented with assumptions about independence. First, a pragmatic one: given a body of default knowledge, the probabilistic interpretation renders a behavior in close correspondence with intuition. This is important as we want the interpretation to be faithful to the information encoded in the knowledge base. The second benefit, is of a more theoretical nature: we can appeal to the empirical grounds of a probabilistic semantics for *understanding* potential disagreements between what is sanctioned and what is intended. This is particularly relevant in scenarios like the "Yale shooting" in which different solutions have often been motivated on different conceptions of where the problem lies.

Actually, the framework for defeasible inference to be developed in this chapter does not handle the Yale shooting scenario properly. The problem is that the

notion of irrelevance used does not take causal considerations into account. Such aspects of default reasoning will be treated in detail in chapter 5.

2.2 Language: Default Theories

The system of defeasible inference to be introduced accepts as input a context composed of sentences and defaults, and implicitly characterizes the conclusions that legitimately follow from that context. Sentences and defaults are expressed in terms of an underlying first order language $\mathcal{L}$. We use the object-level connective '$\Rightarrow$' for material implication, and the meta-level connective '$\rightarrow$' for defaults. The sentence '$p \Rightarrow q$' thus reads as 'if p then q,' while the expression '$p \rightarrow q$' as 'if p, then *normally* q.' The symbols '$\vdash$' and '$\not\vdash$' are used to stand for derivability and non-derivability in classical first order logic with equality, respectively.

We use the letters p, q, ..., possibly indexed, as variables ranging over sentences, and write object-level sentences in typewriter style (e.g. $\texttt{dog(fido)}$). Likewise, letters in italics from the end of the alphabet x, y, ... denote variables (sometimes, tuples of variables), and from the beginning of the alphabet a, b, ... denote ground terms (sometimes, tuples of ground terms). Sentences are implicitly universally quantified, so we often write $\texttt{dog}(x) \Rightarrow \texttt{animal}(x)$, for instance, instead of $\forall x.\, \texttt{dog}(x) \Rightarrow \texttt{animal}(x)$.

Default theories $T = \langle K, E \rangle$ are comprised of two components: a *background context* K containing generic information, and an *evidence set* E containing information specific to the particular situation at hand. Intuitively, K contains the relevant *rules*, while E contains the relevant *facts*. For instance, in the canonical "birds fly, penguins don't" example, we will include the encoding of the defaults "birds fly" and "penguins don't fly" as well as the strict inclusion "penguins are birds," in K, leaving in E facts such as "Tweety is a bird," "Tim weights three pounds," etc. We will also refer to the pair $\langle K, E \rangle$ as a *context* and will sometimes denote it simply as E_K.

The background context $K = \langle L, D \rangle$ is also comprised of two components: a sentential component L and a default component D. While L and E are sets of sentences, namely closed wffs in $\mathcal{L}$, D stands for a set of *defaults*. Defaults are encoded by expressions of the form $p \rightarrow q$, where p and q denote sentences in $\mathcal{L}$ called the default antecedent and consequent, respectively. The expression $\texttt{dog(fido)} \rightarrow \texttt{can\_bark(fido)}$, for instance, represents a default stating that "nor-

mally, if Fido is a dog, Fido can bark." We use *default schemas* of the form $p(x) \rightarrow q(x)$, where p and q are wffs with free variables among those of x, to denote the collection of defaults $p(a) \rightarrow q(a)$ that results from substituting x by all the tuples a of ground terms in the language.

Often it will be convenient to consider default theories in which defaults are associated with *assumptions* in the underlying language $\mathcal{L}$. In such cases we will be able to say that a default is satisfied or is violated in a model according to whether its associated assumption is satisfied or not. For that we will assume a set $\Delta_{\mathcal{L}}$ of assumptions in the language containing all atoms of the form $\delta_i(a)$, where i is some type of index.

An *assumption based default theory* then is a default theory $T = \langle K, E \rangle$ in which the consequent of each default $p \rightarrow \delta$ in K is a unique assumption δ in $\Delta_{\mathcal{L}}$. Arbitrary default theories $T' = \langle K', E \rangle$ can be expressed in an assumption based format by replacing each default schema $p(x) \rightarrow q(x)$ in K' with a sentence $p(x) \wedge \delta_i(x) \Rightarrow q(x)$ and a new default schema $p(x) \rightarrow \delta_i(x)$, for a unique index i. Literals $\delta_i(x)$ provide us with object-level handles on defaults, in a way similar to Poole's [1988] default naming convention, and McCarthy's [1986] abnormality predicates.

A theory with a background context K containing a sentence $\mathtt{bird}(x) \Rightarrow \mathtt{animal}(x)$ and a default schema $\mathtt{bird}(x) \rightarrow \mathtt{fly}(x)$, for example, can be expressed in an assumption based format by replacing the default schema by the expressions $\mathtt{bird}(x) \wedge \delta_1(x) \Rightarrow \mathtt{fly}(x)$ and $\mathtt{bird}(x) \rightarrow \delta_1(x)$, for a unique index 1. We will often abbreviate the later two expressions by simply writing $\mathtt{bird}(x) \rightarrow_1 \mathtt{fly}(x)$.

An important assumption we will adopt in this chapter is that the underlying language $\mathcal{L}$ and the background contexts $K = \langle L, D \rangle$ are such that there is a *finite* number of different truth valuations defined over $\mathcal{L}$ that satisfy the sentences in L. We say in that case that default theories $T = \langle K, E \rangle$ defined over $\mathcal{L}$ are finite. Finite propositional theories as well as default theories augmented by suitable domain closure axioms, for example, will qualify as finite default theories. These conditions will be later relaxed in chapter 4.

2.3 Rules of Inference: The Core

The next system of defeasible inference will be referred to as **P** as an abbreviation for "probabilistic." **P** is characterized by a set of rules of inference in the style of natural deduction systems. The first five rules constitute what we call the *core*. We will later introduce an additional rule of inference which extends the inferential power of the core significantly. The reason to isolate the first five rules as the core of **P**, is because these rules admit a precise and pure probabilistic interpretation. The power and limitations of the core will thus be a good indication of the power and limitations of the underlying probabilistic interpretation. The last rule can be interpreted as supplementing the core with assumptions about independence extracted from the structure of discourse. It will be shown in chapters 3 and 4 that both **P** and its core can be given a *model-theoretic* interpretation as well.

The rules of **P** implicitly define the set of conclusions that follow from a given context. We write $E \mid\!\sim_K p$ to denote that the sentence p is derivable in **P** from a context $T = \langle K, E\rangle$ with background context $K = \langle L, D\rangle$. Likewise, $E, \{q\} \mid\!\sim_K p$, abbreviated $E, q \mid\!\sim_K p$, states that p is derivable from the context that results from adding the sentence q to E. We will use the notation $E \mid\!\sim_K p$ as an abbreviation of $E, L \vdash p$. It should be kept in mind, however, that the consequence operator '$\mid\!\sim_K$,' unlike '$\mid\!\sim_K$,' is *non-monotonic*, so the expression $E, q \mid\!\sim_K p$ does not necessarily follow from $E \mid\!\sim_K p$.

Definition 2.1 *The* core *of* **P** *is defined by the following set of rules:*

Rule 1 (Defaults) *If* $p \rightarrow q \in D$ *then* $p \mid\!\sim_K q$

Rule 2 (Deduction) *If* $E \mid\!\sim_K p$ *then* $E \mid\!\sim_K p$

Rule 3 (Augmentation) *If* $E \mid\!\sim_K p$ *and* $E \mid\!\sim_K q$ *then* $E, p \mid\!\sim_K q$

Rule 4 (Reduction) *If* $E \mid\!\sim_K p$ *and* $E, p \mid\!\sim_K q$ *then* $E \mid\!\sim_K q$

Rule 5 (Disjunction) *If* $E, p \mid\!\sim_K r$ *and* $E, q \mid\!\sim_K r$ *then* $E, p \vee q \mid\!\sim_K r$

The **defaults** rule permits us to conclude the consequent of a default when its antecedent represents all the available evidence. **Deduction** states that whatever the context, what is derivable by the rules of classical logic, is also derivable in **P**. **Augmentation** permits the assimilation of an established conclusion to the current evidence set without affecting the status of any other derived conclusions.

Reduction is the inverse of augmentation: it permits to remove information from the evidence set if such information is derivable from the reduced set. Finally, **disjunction**, permits reasoning by cases.

Rules 2–5 can be shown to share the inferential power of the system proposed by Adams [1966] for deriving what he calls the probabilistic consequences of a given set of conditionals. Some of these rules also appear, in different forms, in several logics of conditionals (see [Nute, 1984]), and in a "minimal" non-monotonic logic proposed by Gabbay [1985]. More recently Gabbay's system has been further investigated by Kraus *et al.* [1988] and Makinson [1989] who arrive at a system which is equivalent to the core above, but which they justify on model-theoretic grounds.[1]

We proceed now to investigate some of the properties of the system defined by rules 1–5. Later on, we will discuss some of its limitations as we enhance the system with an additional inference rule responsible for drawing assumptions about independence.

Some Useful Derived Rules of Inference

The following derived rules of inference illustrate some of the logical properties of **P**:

Theorem 2.1 *The following rules are derived rules of* **P**:

Deductive Closure *If* $E \mid\!\approx_K p$, $E \mid\!\approx_K q$, *and* $E, p, q \mid\!\approx_K r$, *then* $E \mid\!\approx_K r$
Context Equivalence *If* $E, p \mid\!\approx_K q$, $E, q \mid\!\approx_K p$, *and* $E, p \mid\!\approx_K r$, *then* $E, q \mid\!\approx_K r$
Weak Reduction *If* $E, q \mid\!\approx_K p$ *then* $E \mid\!\approx_K \neg q \vee p$
Presuppositions *If* $E \mid\!\approx_K p$ *and* $E, q \mid\!\approx_K \neg p$ *then* $E \mid\!\approx_K \neg q$
Parallel Reduction *If* $E, p, q \mid\!\approx_K r$, $E \mid\!\approx_K p$, *and* $E \mid\!\approx_K q$, *then* $E \mid\!\approx_K r$
OR-transitivity *If* $E, p \vee q \mid\!\approx_K q$ *and* $E, q \vee r \mid\!\approx_K r$, *then* $E, p \vee r \mid\!\approx_K r$
OR-monotonicity *If* $E, p \vee q \mid\!\approx_K \neg q$, *then* $E, p \vee q \vee r \mid\!\approx_K \neg q$

To illustrate how derivations proceed in **P**, we include here the corresponding proofs.

[1]More about this in chapter 3.

Proof We start with **deductive closure**. From $E \mathrel{\sim_K} q$ and $E \mathrel{\sim_K} p$, we can obtain $E, p \mathrel{\sim_K} q$ by augmentation. Similarly, from $E, p, q \mathrel{\sim_K} r$ we get $E, p, q \mathrel{\sim_K} r$ by deduction. Applying reduction twice then, the target result follows. For **context-equivalence**, $E, p \mathrel{\sim_K} q$ permits us to augment $E, p \mathrel{\sim_K} r$ into $E, p, q \mathrel{\sim_K} r$, while $E, q \mathrel{\sim_K} p$ permits us to reduce the latter into the desired conclusion $E, q \mathrel{\sim_K} r$. For **Weak reduction** note that $E, \neg q \mathrel{\sim_K} \neg q \vee p$ and $E, q \mathrel{\sim_K} \neg q \vee p$, follow by deduction and deductive closure respectively. Thus, reasoning by cases and the reducing $q \vee \neg q$ from the evidence set, the final result is obtained. **Presupposition** is a consequence of weak reduction on $E, q \mathrel{\sim_K} \neg p$ and the deductive closure of $E \mathrel{\sim_K} \neg q \vee \neg p$ and $E \mathrel{\sim_K} p$. **Parallel reduction** follows from the augmentation of $E \mathrel{\sim_K} p$ into $E, q \mathrel{\sim_K} p$, and the reduction of $E, q, p \mathrel{\sim_K} r$ into $E, q \mathrel{\sim_K} r$, and further into $E \mathrel{\sim_K} r$. To prove **Or-transitivity**, note that by deduction we can obtain $E, q \vee r \mathrel{\sim_K} p \vee q \vee r$, while from the hypothesis $E, p \vee q \mathrel{\sim_K} q$ and reasoning by cases we get $E, p \vee q \vee r \mathrel{\sim_K} q \vee r$. Similarly, we obtain $E, p \vee r \mathrel{\sim_K} p \vee q \vee r$ and $E, p \vee q \vee r \mathrel{\sim_K} p \vee r$. Finally, from the hypothesis $E, q \vee r \mathrel{\sim_K} r$ one application of context-equivalence yields $E, p \vee q \vee r \mathrel{\sim_K} r$, and a second $E, p \vee r \mathrel{\sim_K} r$. **OR-monotonicity** is a consequence of augmenting the hypothesis $E, p \vee q \mathrel{\sim_K} \neg q$ into $E, p \vee q, p \vee q \vee r \mathrel{\sim_K} \neg q$, from which the conclusion $E, p \vee q \vee r \mathrel{\sim_K} \neg q$ follows by weakly reducing $p \vee q$ from the premises, and deductive closure.

Some non-theorems:

$E \vdash p$ and $p \mathrel{\sim_K} q$ do necessarily imply $E \mathrel{\sim_K} q$
$E \mathrel{\sim_K} p$ and $E' \mathrel{\sim_K} p$ do not necessarily imply $E, E' \mathrel{\sim_K} p$

Note that the first non-theorem is clearly undesirable. If accepted, it would endow our system with the monotonic characteristics of classical logic, precluding exceptions like non-flying birds, etc. The second one would authorize conclusions such that John will be happy when married to both Jane and Mary on the grounds that he will be happy when married to either one of them.

The system of rules 1–5 defines an extremely conservative non-monotonic logic. In fact, the inferences sanctioned by these rules do not invoke any assumptions regarding information absent from the background context. Namely, while the core is *non-monotonic* in the evidence set E, it is *monotonic* in the background context K. If for two background contexts $K = \langle L, D \rangle$ and $K' = \langle L', D' \rangle$ we write $K \subseteq K'$ for $L \subseteq L'$ and $D \subseteq D'$, the following theorem holds:[2]

[2]Proofs can be found in the appendix.

Theorem 2.2 (K-monotonicity) *If* $E \mathrel{\mid\!\!\sim_K} p$ *and* $K \subseteq K'$ *then* $E \mathrel{\mid\!\!\sim_{K'}} p$

2.4 Semantics: ϵ-entailment

As indicated above, it is possible to construct a probabilistic interpretation under which rules 1–5 can be shown to be *sound* and, as we will later see, *complete*. The idea, roughly, is to associate the expression $E \mathrel{\mid\!\!\sim_K} p$ with conditional probability statements $P_K(p \mid E) \approx 1$, for probability distributions $P_K(\cdot)$ which comply with the constraints in K in a suitable way. We provide the details of such an interpretation below. Note, however, the different roles that *background* and *evidence* will play: while the background K delimits the space of probability distributions to be considered to what we call the *admissible* probability distributions $P_K(\cdot)$; the set E participates as the information on which the admissible probability distributions are *conditioned* upon.[3] More precisely, the admissibility of a probability distribution P relative to a background context K and a given *range* ϵ is defined as follows.

Definition 2.2 *A probability distribution* P_K *is* admissible *with a background context* $K = \langle L, D \rangle$ within a range ϵ, *if* P_K *assigns unit probability to every sentence* s *in* L, *i.e.* $P_K(s) = 1$, *and for each default* $p \rightarrow q$ *in* D, $P_K(q \mid p)$ *is greater than* $1 - \epsilon$, *while* $P_K(p)$ *is greater than zero.*

In other words, a probability distribution is admissible within a range ϵ when it renders the sentences in L *certain*, while leaving a range ϵ of uncertainty for the defaults in D. What we show next is a result due to Adams [1966, 1975], stating that when the expression $E \mathrel{\mid\!\!\sim_K} p$ is derivable by means of rules 1–5, the conditional probability $P_K(p \mid E)$ is bound to approach to one, as the range of uncertainty ϵ approaches zero. When this happens, we say that the proposition p is ϵ-*entailed* by the default theory $T = \langle K, E \rangle$.[4]

Definition 2.3 *A proposition* p *is* ϵ-entailed *by a default theory* $T = \langle K, E \rangle$ *when for any* $\epsilon > 0$, *there is an* $\epsilon' > 0$, *such that for any probability distribution* P_K

[3] The reader without a basic background in probabilities will found in [Pearl, 1988b] all what we are going to need and considerably more.

[4] The terms ϵ-entailment and ϵ-semantics were coined by Pearl in [Pearl and Geffner, 1988]. Adams [1966, 1975], refers to the same notions as probabilistic entailment (p-entailment) and probabilistic semantics, respectively. Here we will adhere to the ϵ-terminology, leaving the term p-entailment for other purposes.

admissible with K within a range ϵ', $P_K(p \mid E) > 1 - \epsilon$.

The soundness result can then be cast as follows:

Theorem 2.3 (Adams) *If $E \mathrel{|\!\sim_K} p$ then p is ϵ-entailed by the default theory $T = \langle K, E \rangle$.*

Like deductive inference preserves truth, rules 1–5 preserve high-probability. However, unlike classical model-theory, the probabilistic interpretation provides a semantics for default *inference* as opposed to defaults themselves. We cannot evaluate whether a default is true in given a world; rather, the interpretation tell us which propositions should be accepted, provided (1) that only conclusions with arbitrarily high conditional probability (short of one) are accepted, and (2) that defaults are accepted.[5]

A distinctive feature of this interpretation is that it leads to a notion of *default consistency* which is different from traditional ones. A background context K containing two defaults $p \rightarrow q$ and $p \rightarrow \neg q$, for instance, legitimizes any sentence in the language from a context which includes p. Indeed, inconsistencies may occur when K does not accept admissible probability distributions beyond certain ranges:

Definition 2.4 (Consistency) *A background context $K = \langle L, D \rangle$ is ϵ-consistent, iff there is a probability distribution admissible with K within any positive range. Otherwise, K is ϵ-inconsistent.*

A context $T = \langle K, E \rangle$ whose background context $K = \langle L, D \rangle$ is ϵ-consistent, is thus guaranteed not to sanction logically incompatible propositions as long as the set $L + E$ of sentences is logically consistent. Moreover, as noted by Adams [1975], ϵ-consistency can be determined in terms of ϵ-entailment, and vice versa:

Theorem 2.4 (Adams) *The proposition q is ϵ-entailed by the default theory $T = \langle K, \{p\} \rangle$, with $K = \langle L, D \rangle$ if and only if the background $K' = \langle L, D + \{p \rightarrow \neg q\} \rangle$ is ϵ-inconsistent.*

[5]For alternative probabilistic semantics of plausible reasoning, see Wellman [1987], Neufeld and Poole [1988], Bacchus [1989], and the recent survey by Pearl [1989a].

Such a correspondence between ϵ-entailment and ϵ-consistency will be at the center of the equivalence between ϵ-entailment and a model-theoretic form of defeasible entailment to be studied in the next chapter. We will also show then that the core of **P**, comprised by rules 1–5 above, is not only sound with respect to ϵ-entailment, but in a suitable sense, also complete.

The example below illustrates some of the virtues and limitations of the core as a system of default inference.

Example 2.1 (Specificity) Let us consider a background context K with information about birds (B), red birds (RB), penguins (P) and flying things (F) expressed in an assumption based default theory with formulas (see fig. 2.1):

$$\begin{array}{ll} \mathtt{P}(x) \Rightarrow \mathtt{B}(x) & \mathtt{RB}(x) \Rightarrow \mathtt{B}(x) \\ \mathtt{B}(x) \wedge \delta_1(x) \Rightarrow \mathtt{F}(x) & \mathtt{B}(x) \rightarrow \delta_1(x) \\ \mathtt{P}(x) \wedge \delta_2(x) \Rightarrow \neg\mathtt{F}(x) & \mathtt{P}(x) \rightarrow \delta_2(x) \end{array}$$

Recall that the last two rows encode the defaults $\mathtt{B}(x) \rightarrow \mathtt{F}(x)$ ("birds fly") and $\mathtt{P}(x) \rightarrow \neg\mathtt{F}(x)$ ("penguins don' fly"). The convenience of such an encoding device will become apparent in the next section.

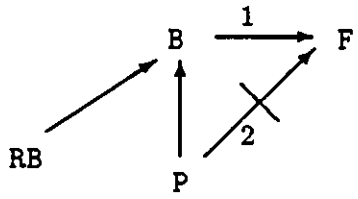


Figure 2.1: The canonical example: birds and penguins

In this background context K, it is possible to prove that an arbitrary bird, say tim, is likely to fly, as follows:

1. $\mathtt{B}(\mathtt{tim}) \mathrel{\tilde{\vdash}}_K \delta_1(\mathtt{tim})$; Defaults $\mathtt{B}(x) \rightarrow \delta_1(x)$
2. $\mathtt{B}(\mathtt{tim}) \mathrel{\tilde{\vdash}}_K \mathtt{F}(\mathtt{tim})$; Deductive Closure 1

The fact that the consequence relation ' $\mid\!\sim_K$ ' is closed under deduction permits us to obtain the behavior associated with the intended default $\mathtt{B(tim)} \rightarrow \mathtt{F(tim)}$ from the encoding in terms of the assumption $\delta_1(\mathtt{tim})$. Indeed, the pattern of 'applying' a default such as $\mathtt{B(tim)} \rightarrow \delta_1(\mathtt{tim})$ and closing the result under deduction will be so common that we will find useful to replace it by the application of the "virtual" default $\mathtt{B(tim)} \rightarrow_1 \mathtt{F(tim)}$ in K. So, we justify the inference that an arbitrary penguin does not fly as:

3. $\mathtt{P(tim)} \mid\!\sim_K \neg\mathtt{F(tim)}$; Defaults $\mathtt{P}(x) \rightarrow_2 \neg\mathtt{F}(x)$

Notice that there is no contradiction between the conclusions $\mathtt{B(tim)} \mid\!\sim_K \mathtt{F(tim)}$ and $\mathtt{P(tim)} \mid\!\sim_K \neg\mathtt{F(tim)}$ as both refer to different contexts: $\{\mathtt{B(tim)}\}_K$ and $\{\mathtt{P(tim)}\}_K$. Nonetheless, since penguins are known to be birds, the latter context subsumes the former one:

4. $\mathtt{P(tim)} \mid\!\sim_K \mathtt{B(tim)}$; Deduction

enabling us to augment 3 above to yield:

5. $\mathtt{P(tim)}, \mathtt{B(tim)} \mid\!\sim_K \neg\mathtt{F(tim)}$; Augmentation 3,4

Thus we see that subclasses properties override classes properties. However, unlike default logic or circumscription, the 'abnormality' of subclasses does not need to be specificed explicitly; the expected behavior emerges automatically from the probabilistic interpretation embodied in the rules and the distinction made between formulas in the background context K from those in the evidence set E. Indeed the behavior we have just illustrated would not be sanctioned if we had included the fact that 'penguins are birds' in E rather than in K. In such a case we would need to show that the expression $\mathtt{P}(x) \Rightarrow \mathtt{B}(x)$ can be assimilated into the left hand side of the non-monotonic consequence operator without affecting the status of the intended conclusion $\neg\mathtt{F(tim)}$. When the expression $\mathtt{P}(x) \Rightarrow \mathtt{B}(x)$ is included in the background context, on the other hand, it does not need to be assimilated; defaults $p \rightarrow q$ in K are interpreted by rule 1 as stating that "if p then q, *even if* K." In particular, thus, if K were augmented with the fact that tim is an exceptional bird, namely $\mathtt{B(tim)} \wedge \neg\mathtt{F(tim)}$, K would become *ϵ-inconsistent:* no probability distribution would be able to make the probability of $\mathtt{F(tim)}$ given $\mathtt{B(tim)}$ arbitrarily high, while making the probability of $\mathtt{B(tim)} \wedge \neg\mathtt{F(tim)}$ one and

the probability of B(tim) greater than zero. That is the reason why the background context is to contain only *generic information*, leaving the information specific to the situation at hand in the evidence set.

While the interpretation of defaults embedded in rules 1–5 captures certain pattern of reasoning that escape more traditional formulations, it falls short in other important aspects. For instance, in the example above, we also would expect that a *red* bird is likely to fly. However, while the expression $\mathtt{B(tim)} \mid\!\sim_K \mathtt{F(tim)}$ ("a bird flies") is derivable, the expression $\mathtt{RB(tim)}, \mathtt{B(tim)} \mid\!\sim_K \mathtt{F(tim)}$ ("a *red* bird flies") is not.

This limitation of rules 1–5 is indeed serious and prevents us from maintaining a conclusion in the presence of additional but *irrelevant* information. Such a behavior, however, is not surprising given the probabilistic semantics underlying the core. Rules 1–5 are probabilistically sound, and therefore, only 'jump' to conclusions $E \mid\!\sim_K p$ whose high conditional probability $P_K(p \mid E)$ can be guaranteed in *every* probability distribution P_K admissible with K.[6] Since there are probability distributions admissible with K in which red birds do not fly (as much as penguins do not fly), the conclusion that a red bird is likely to fly is *not* probabilistically sound, and therefore, not derivable from the core.

To account for these inferences, additional restrictions on the space of probability distributions considered are needed. A natural restriction is to require these distributions to comply with certain assumptions about conditional independence; namely, that the consequent q of a default $p \rightarrow q$ be derivable from its antecedent, as long the available evidence does not indicate otherwise. In the case of assumption based default theories, where defaults are associated with assumptions, such independence assumptions take a form familiar to other non-monotonic formalisms: *assumptions* will be adopted in the absence of conflicting evidence. In the example above, this amounts to maintain the assumptions $\delta_1(\mathtt{tim})$ ("if Tim is a bird, Tim flies") in the presence of the new information RB(tim) ("Tim is a red bird"), as the latter does not provide an argument against the former.

[6]The expression 'a probability distribution admissible with K' is to be understood as 'a probability distribution admissible with K *within infinitesimal ranges*'.

2.5 Independence Assumptions

The account of the irrelevance conditions determines whether an assumption δ can be asserted in a given context $T = \langle K, E \rangle$ by checking whether T provides *arguments* against δ. Probabilistically, if δ is the assumption associated with a default $p \rightarrow q$ in K, such conditions amount checking whether the sentence $p \Rightarrow q$ can be assumed to be independent of the body of evidence E in K; namely, if $P_K(p \Rightarrow q \,|\, E)$ can be assumed to be equal to $P_K(p \Rightarrow q)$ for any admissible probability distribution P_K.[7]

First, we shall make the notion of *arguments* precise. We use the symbols Δ, Δ', ... to stand for sets of assumptions, and $\neg\Delta$ to stand for the negation of the conjoin of Δ ($\neg\Delta \equiv$ **true** if Δ is empty).

Definition 2.5 *A set of assumptions* Δ *constitutes an* argument *in a context* $T = \langle K, E \rangle$, *iff* $E \not\vdash_K \neg\Delta$. *An argument* Δ *is an argument* for *a proposition* p *in* T, *iff* $E, \Delta \vdash_K p$, *and an argument* against p *iff* $E, \Delta \vdash_K \neg p$.

Intuitively, if a body of evidence E does not give rise to arguments against an assumption δ, it is reasonable to assume that E is *irrelevant* to the status of δ. In the example above, this amounts to say that given the evidence $E = \{\mathtt{B(tim)}, \mathtt{RB(tim)}\}$, the assumption $\delta_1(\mathtt{tim})$ associated with the default $\mathtt{B(tim)} \rightarrow_1 \mathtt{F(tim)}$ can be assumed to hold, enabling a derivation for $\mathtt{B(tim)}, \mathtt{RB(tim)} \mathrel{|\!\sim_K} \mathtt{F(tim)}$ by deductive closure.

The problem, however, is that the characterization of the irrelevance of a body of evidence E to an assumption δ in terms of the lack of counterarguments is too weak. Many times it is possible to construct arguments against an assumption which, intuitively, do not count. In the same example above, this happens when all we know about Tim is that it is a penguin, i.e. $E = \{\mathtt{P(tim)}\}$. In such a context, an argument $\{\delta_1(\mathtt{tim})\}$ ("if Tim is a bird, Tim flies") against the assumption $\delta_2(\mathtt{tim})$ ("if Tim is a penguin, Tim does not fly") can be constructed, and still $\delta_2(\mathtt{tim})$ is derivable from $T = \langle K, E \rangle$ by means of rule 1. In other words, the argument $\{\delta_1(\mathtt{tim})\}$ refutes the assumption $\delta_2(\mathtt{tim})$ but carries no weight. However, if we want to assess the relevance of an additional piece of evidence, say $\mathtt{RB(tim)}$, such an

[7] The account in [Geffner and Pearl, 1987] and [Geffner, 1988] computes whether the condition $P_K(q \,|\, p, E) \approx P_K(q \,|\, p)$ can be assumed to hold, which unlike the conditions addressed here, does not allow for default contraposition.

argument remains, preventing to tag the evidence $E' = \{\mathtt{P(tim)}, \mathtt{RB(tim)}\}$ as irrelevant to $\delta_2(\mathtt{tim})$, and thus precluding a derivation for $\mathtt{P(tim)}, \mathtt{RB(tim)} \mid\!\!\sim_K \neg\mathtt{F(tim)}$.

What such a scenario suggests is that the license to derive an assumption δ from p in a background K containing a default $p \rightarrow \delta$ (rule 1), presumes that arguments Δ against δ in the context $T = \langle K, \{p\}\rangle$ carry no weight. Let us say that such arguments are against the *default* $p \rightarrow \delta$:

Definition 2.6 *A set of assumptions Δ is an argument against a* default $p \rightarrow \delta$ *in K, iff Δ is an argument against the* assumption δ *in the context $T = \langle K, \{p\}\rangle$.*

The intuitions then is that when assessing the relevance of E to δ, arguments Δ against an assumption δ which are also arguments against a default $p \rightarrow \delta$ in K should not be considered. We say that Δ is an argument which is *directly dominated* by δ:

Definition 2.7 (Direct Dominance) *A set of assumptions Δ is* directly dominated (d-dominated) *by an assumption δ in K, iff Δ contains an argument Δ' against a default $p \rightarrow \delta$ in K.*

It is easy to check that the expected behavior follows from the example above once dominated counterarguments are discarded. Indeed, the evidence $E' = \{\mathtt{P(tim)}, \mathtt{RB(tim)}\}$, gives rise to a single minimal argument $\{\delta_1(\mathtt{tim})\}$ ("if Tim is a bird, then Tim flies") against the assumption $\delta_2(\mathtt{tim})$ which is d-dominated by $\delta_2(\mathtt{tim})$. Thus, if such argument is ignored, the expression $\mathtt{P(tim)}, \mathtt{RB(tim)} \mid\!\!\sim_K \delta_2(\mathtt{tim})$, and by means of deductive closure, the expression $\mathtt{P(tim)}, \mathtt{RB(tim)} \mid\!\!\sim_K \neg\mathtt{F(tim)}$ ("a red penguin doesn't fly") would be authorized.

This irrelevance criterion is sufficient for many other cases as well, and moreover, can be shown not to introduce any inconsistencies (chapter 4). However, it is not strong enough. Consider for instance the theory that results from the example above by replacing the *strict* inclusion $\mathtt{P}(x) \Rightarrow \mathtt{B}(x)$ by a *default* inclusion $\mathtt{P}(x) \rightarrow_3 \mathtt{B}(x)$. In the resulting background context, $\{\delta_1(\mathtt{tim})\}$ is no longer an argument against the default $\mathtt{P(tim)} \rightarrow \delta_2(\mathtt{tim})$, and thus it is no longer d-dominated by the assumption $\delta_2(\mathtt{tim})$. As a result, the body of evidence $E' = \{\mathtt{P(tim)}, \mathtt{RB(tim)}\}$ can no longer be proven irrelevant to $\delta_2(\mathtt{tim})$, precluding thus a derivation for $\mathtt{P(tim)}, \mathtt{RB(tim)} \mid\!\!\sim_K \neg\mathtt{F(tim)}$.

Note, however, that the assumption $\delta_2(\texttt{tim})$ directly dominates the set of assumptions $\{\delta_1(\texttt{tim}), \delta_3(\texttt{tim})\}$, and the assumption $\delta_3(\texttt{tim})$ directly dominates the set $\{\delta_1(\texttt{tim}), \delta_2(\texttt{tim})\}$. Thus, *if the relation "δ directly dominates the set Δ" is understood as meaning that the assumption δ has a higher priority than some assumption in Δ,* where priority relations are *irreflexive* and *transitive*, both $\delta_2(\texttt{tim})$ and $\delta_3(\texttt{tim})$ are forced to have a higher priority than $\delta_1(\texttt{tim})$. Intuitively, then, the argument $\{\delta_1(\texttt{tim})\}$ should still be dominated by the assumption $\delta_2(\texttt{tim})$, and thus ignored when assessing the relevance of E' to $\delta_2(\texttt{tim})$.

The following extended definition of *dominance* captures such understanding of direct dominance in terms of priorities:

Definition 2.8 (Dominance) *The* set Δ *of assumptions* dominates *a set Δ' relative to a background context K, iff every assumption δ in Δ directly dominates the set $\Delta + \Delta'$. An* assumption δ dominates *a set Δ' iff δ belongs to a set Δ that dominates Δ'.*

In chapter 4 we will analyze such a prioritized interpretation of defaults in detail, and prove that δ must have a higher priority than some assumption in each set it *dominates,* provided that every assumption has a higher priority than some assumption in each set it *directly dominates.*

Since arguments against δ which rely on assumptions with lower priority than δ should be discounted, the following definition of the *irrelevance* conditions finally results:

Definition 2.9 (Irrelevance) *A body of evidence E is* irrelevant *to an assumption δ in a background context K, written $I_K(\delta \mid E)$, iff every argument against δ in the context $T = \langle E, K\rangle$ is* dominated *by δ.*

The last inference rule of the system **P** permits us to derive an assumption δ in a context $T = \langle K, E\rangle$ provided that the body of evidence E is irrelevant to δ:

Rule 6 (Irrelevance) If $I_K(\delta \mid E)$ then $E \mathrel{\sim_K} \delta$

As a special case, for an assumption δ_i associated with a default $p \rightarrow_i q$ in K, we will often find useful to write $I_K(p \rightarrow_i q \mid E)$ as an abbreviation of $I_K(\delta_i \mid E + \{p\})$. We will say in that case that E is irrelevant to the *default* $p \rightarrow_i q$, and invoke the irrelevance rule in the following, more restricted form:

Rule 6.1 (Irrelevance) If $I_K(p \rightarrow_i q \,|\, E)$ then $E, p \mathrel{|\!\sim_K} q$

Provided with this irrelevance rule, a proof that a red bird is expected to fly (example 2.1), can then be constructed as follows:

1. $I_K(\mathtt{B(tim)} \rightarrow_1 \mathtt{F(tim)} \,|\, \mathtt{RB(tim)})$; No arguments against $\delta_1(\mathtt{tim})$
2. $\mathtt{RB(tim)}, \mathtt{B(tim)} \mathrel{|\!\sim_K} \mathtt{F(tim)}$; Irrelevance (6.1) 1
3. $\mathtt{RB(tim)} \mathrel{|\!\sim_K} \mathtt{B(tim)}$; Deduction
4. $\mathtt{RB(tim)} \mathrel{|\!\sim_K} \mathtt{F(tim)}$; Reduction 2,3.

In chapter 4 a justification for the irrelevance rule in terms of priorities will be considered in detail.

2.6 Examples

In this section we illustrate the behavior of the system of defeasible inference determined by rules 1–6 on a number of examples.

Example 2.2 (Default Preferences) Let the background context K contain the following defaults : "adults (A) work (W)","university students (U) are adults but do not work", and "adults which are young (Y) are university students", expressed as:

$$\mathtt{A}(x) \rightarrow_1 \mathtt{W}(x)$$
$$\mathtt{U}(x) \rightarrow_2 \mathtt{A}(x)$$
$$\mathtt{U}(x) \rightarrow_3 \neg\mathtt{W}(x)$$
$$\mathtt{A}(x) \wedge \mathtt{Y}(x) \rightarrow_4 \mathtt{U}(x)$$

Namely, for each expression $p(x) \rightarrow_i q(x)$, K contains a default schema $p(x) \rightarrow \delta_i(x)$ and a sentence $p(x) \wedge \delta_i(x) \Rightarrow q(x)$. Figure 2.2 provides a graphical representation of K. The labels on links indicate the indices of the associated assumption predicates.

We show first that an adult, say Ken (k), who is also a university student, is likely not to work, i.e. $\mathtt{A(k)}, \mathtt{U(k)} \mathrel{|\!\sim_K} \neg\mathtt{W(k)}$:

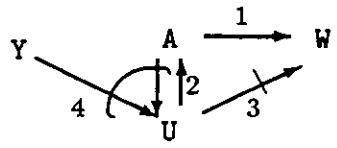


Figure 2.2: Implicit preferences among defaults

1. $\mathtt{U(k)} \mid\!\sim_K \neg\mathtt{W(k)}$; Defaults, $\mathtt{U}(x) \rightarrow_3 \mathtt{W}(x)$
2. $\mathtt{U(k)} \mid\!\sim_K \mathtt{A(k)}$; Defaults, $\mathtt{U}(x) \rightarrow_2 \mathtt{A}(x)$
3. $\mathtt{U(k)}, \mathtt{A(k)} \mid\!\sim_K \neg\mathtt{W(k)}$; Augmentation 1,2

Note that the rules in **P** result again in a *preference* for more 'specific' defaults. Indeed, if we consider the context $\{\mathtt{U(k)}\}_K$ it is possible to find a *preference* pattern among assumptions. Namely, while the assumptions $\delta_1(\mathtt{k})$ and $\delta_2(\mathtt{k})$ and $\delta_3(\mathtt{k})$ are in conflict, $\delta_2(\mathtt{k})$ and $\delta_3(\mathtt{k})$ hold, but $\delta_1(\mathtt{k})$ does not:

4. $\mathtt{U(k)} \mid\!\sim_K \delta_2(\mathtt{k}) \wedge \delta_3(\mathtt{k}) \wedge \neg\delta_1(\mathtt{k})$; Defaults + Deductive Closure

Indeed, the set of assumptions $\Delta = \{\delta_2(\mathtt{k}), \delta_3(\mathtt{k})\}$ *dominates* the set of assumptions $\Delta' = \{\delta_1(\mathtt{k})\}$, as each assumption in Δ directly dominates the set $\Delta + \Delta'$:

$$\mathtt{U(k)}, \delta_1(\mathtt{k}), \delta_3(\mathtt{k}) \mid\!\sim_K \neg\delta_2(\mathtt{k}) \quad \text{and} \quad \mathtt{U(k)} \not\mid\!\sim_K \neg(\delta_1(\mathtt{k}) \wedge \delta_3(\mathtt{k}))$$
$$\mathtt{U(k)}, \delta_1(\mathtt{k}), \delta_2(\mathtt{k}) \mid\!\sim_K \neg\delta_3(\mathtt{k}) \quad \text{and} \quad \mathtt{U(k)} \not\mid\!\sim_K \neg(\delta_1(\mathtt{k}) \wedge \delta_2(\mathtt{k}))$$

In light of the definitions of irrelevance and dominance above, this implies that arguments against either $\delta_2(\mathtt{k})$ or $\delta_3(\mathtt{k})$ which appeal to the assumption $\delta_1(\mathtt{k})$ can be ignored. In particular, then, it is possible to construct an alternative derivation for $\mathtt{U(k)}, \mathtt{A(k)} \mid\!\sim_K \neg\mathtt{W(k)}$ which relies on the irrelevance of the evidence $E = \{\mathtt{U(k)}, \mathtt{A(k)}\}$ to the status of the assumption $\delta_3(\mathtt{k})$:

5. $I_K(\delta_3(\mathtt{k}) \mid \mathtt{U(k)}, \mathtt{A(k)})$; Defn + $\{\delta_2(\mathtt{k}), \delta_3(\mathtt{k})\}$ dominates $\{\delta_1(\mathtt{k})\}$
6. $\mathtt{U(k)}, \mathtt{A(k)} \mid\!\sim_K \delta_3(\mathtt{k})$; Irrelevance 5
7. $\mathtt{U(k)}, \mathtt{A(k)} \mid\!\sim_K \neg\mathtt{W(k)}$; Deductive Closure 6

The irrelevance conditions, however, allow us to go well beyond what is derivable by the core. In particular, the expression $\mathtt{A(k)}, \mathtt{Y(k)}, \mathtt{U(k)} \mid\!\sim_K \neg\mathtt{W(k)}$, while not derivable by means of rules 1–5, has a derivation in the extended system. Such a derivation indeed is equivalent to steps 5–7 above, except that the set $\{\mathtt{A(k)}, \mathtt{U(k)}\}$ needs to be replaced by the set $\{\mathtt{A(k)}, \mathtt{Y(k)}, \mathtt{U(k)}\}$. Another expression not derivable

from the core alone is $\mathtt{A(k)}, \mathtt{Y(k)} \mathrel{|\!\sim_K} \neg\mathtt{W(k)}$; namely, that a young adult is likely not to work. The irrelevance rule permits the following derivation:

8.	$I_K(\delta_4(\mathtt{k}) \mid \mathtt{A(k)}, \mathtt{Y(k)})$	; Defn + $\{\delta_2(\mathtt{k}), \delta_3(\mathtt{k}), \delta_4(\mathtt{k})\}$ dominates $\{\delta_1(\mathtt{k})\}$
9.	$I_K(\delta_3(\mathtt{k}) \mid \mathtt{A(k)}, \mathtt{Y(k)})$	; Def + $\{\delta_2(\mathtt{k}), \delta_3(\mathtt{k})\}$ dominates $\{\delta_1(\mathtt{k})\}$
10.	$\mathtt{A(k)}, \mathtt{Y(k)} \mathrel{|\!\sim_K} \delta_4(\mathtt{k})$	; Irrelevance 8
11.	$\mathtt{A(k)}, \mathtt{Y(k)} \mathrel{|\!\sim_K} \delta_3(\mathtt{k})$	; Irrelevance 9
12.	$\mathtt{A(k)}, \mathtt{Y(k)} \mathrel{|\!\sim_K} \neg\mathtt{W(k)}$	; Deductive Closure 10, 11

We will later show, that while the irrelevance rule permits us to go beyond the core, it is guaranteed not to lead to inconsistencies (chapter 4). For instance, there is no way to prove the conclusion $\mathtt{A(k)}, \mathtt{Y(k)} \mathrel{|\!\sim_K} \mathtt{W(k)}$; the evidence $E = \{\mathtt{A(k)}, \mathtt{Y(k)}\}$ is indeed *relevant* to the assumption $\delta_1(\mathtt{k})$ by means of the counter-argument $\Delta = \{\delta_2(\mathtt{k}), \delta_3(\mathtt{k}), \delta_4(\mathtt{k})\}$ which is not dominated by $\delta_1(\mathtt{k})$.

It is also possible to illustrate in this example, how default contraposition works in **P**. For instance, we can derive that if Ken does not work he is likely not be an adult:

13.	$I_K(\delta_1(\mathtt{k}) \mid \neg\mathtt{W(k)})$	; Definition + No counterarguments
14.	$\neg\mathtt{W(k)} \mathrel{|\!\sim_K} \delta_1(\mathtt{k})$	; Irrelevance 13
15.	$\neg\mathtt{W(k)} \mathrel{|\!\sim_K} \neg\mathtt{A(k)}$	; Deductive Closure 14

However, if it is additionally learned that Ken is a university student, the conclusion changes; and while the former derivation no longer applies, the following one does:

16.	$\mathtt{U(k)} \mathrel{|\!\sim_K} \mathtt{A(k)}$	; Defaults $\mathtt{U}(x) \rightarrow_2 \mathtt{A}(x)$
17.	$\mathtt{U(k)} \mathrel{|\!\sim_K} \neg\mathtt{W(k)}$	; Defaults $\mathtt{U}(x) \rightarrow_3 \neg\mathtt{W}(x)$
18.	$\mathtt{U(k)}, \neg\mathtt{W(k)} \mathrel{|\!\sim_K} \mathtt{A(k)}$	; Augmentation 16,17

The reason the former derivation does not hold in the new context is that the irrelevance assertion $I_K(\delta_1(\mathtt{k}) \mid \mathtt{U(k)}, \neg\mathtt{W(k)})$ is false: there is an argument $\{\delta_2(\mathtt{k})\}$ against the assumption $\delta_1(\mathtt{k})$ in the context $\{\mathtt{U(k)}, \neg\mathtt{W(k)}\}_K$ which is not dominated by $\delta_1(\mathtt{k})$.

Example 2.3 (Cases) Let us consider now a background context K with defaults "quakers (q) are doves (d)," "republicans (r) are hawks (h)" and "both doves and hawks are politically motivated (p)," together with the fact that nobody is both a

hawk and a dove:[8]

$$\begin{aligned} \mathtt{r}(x) &\rightarrow_1 \mathtt{h}(x) \\ \mathtt{q}(x) &\rightarrow_2 \mathtt{d}(x) \\ \mathtt{h}(x) &\rightarrow_3 \mathtt{p}(x) \\ \mathtt{d}(x) &\rightarrow_4 \mathtt{p}(x) \\ \neg(\mathtt{d}(x) &\wedge \mathtt{h}(x)) \end{aligned}$$

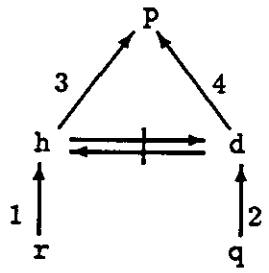


Figure 2.3: Reasoning by cases

We want to show that somebody, say Nixon (n), who is both a quaker and a republican is likely to be politically motivated. The proof involves first reasoning by cases to get $\mathtt{q}(\mathtt{n}), \mathtt{r}(\mathtt{n}) \mathrel{|\!\sim_K} \mathtt{d}(\mathtt{n}) \vee \mathtt{h}(\mathtt{n})$. The first case, $\mathtt{q}(\mathtt{n}), \mathtt{r}(\mathtt{n}), \mathtt{d}(\mathtt{n}) \mathrel{|\!\sim_K} \mathtt{d}(\mathtt{n}) \vee \mathtt{h}(\mathtt{n})$ trivially follows by deduction. The second case, $\mathtt{q}(\mathtt{n}), \mathtt{r}(\mathtt{n}), \neg\mathtt{d}(\mathtt{n}) \mathrel{|\!\sim_K} \mathtt{d}(\mathtt{n}) \vee \mathtt{h}(\mathtt{n})$, requires the irrelevance of $E = \{\mathtt{q}(\mathtt{n}), \neg\mathtt{d}(\mathtt{n})\}$ to $\mathtt{r}(\mathtt{n}) \rightarrow_1 \mathtt{h}(\mathtt{n})$. This is straightforward, as there are no arguments against the assumption $\delta_1(\mathtt{n})$ in the resulting context. Finally, since by irrelevance we can further conclude $\mathtt{q}(\mathtt{n}), \mathtt{r}(\mathtt{n}) \mathrel{|\!\sim_K} \delta_3(\mathtt{n})$ and $\mathtt{q}(\mathtt{n}), \mathtt{r}(\mathtt{n}) \mathrel{|\!\sim_K} \delta_4(\mathtt{n})$, the target conclusion $\mathtt{q}(\mathtt{n}), \mathtt{r}(\mathtt{n}) \mathrel{|\!\sim_K} \mathtt{p}(\mathtt{n})$ follows by deductive closure. Note, however, that neither $\mathtt{q}(\mathtt{n}), \mathtt{r}(\mathtt{n}) \mathrel{|\!\sim_K} \mathtt{d}(\mathtt{n})$ nor $\mathtt{q}(\mathtt{n}), \mathtt{r}(\mathtt{n}) \mathrel{|\!\sim_K} \mathtt{h}(\mathtt{n})$ are sanctioned by $\mathbf{P}$, as the required irrelevance conditions $I_K(\mathtt{r}(\mathtt{n}) \rightarrow_1 \mathtt{h}(\mathtt{n}) \mid \mathtt{q}(\mathtt{n}))$ and $I_K(\mathtt{q}(\mathtt{n}) \rightarrow_2 \mathtt{d}(\mathtt{n}) \mid \mathtt{r}(\mathtt{n}))$ do not hold.

Example 2.4 (Inconsistency) Defaults in $\mathbf{P}$ may give rise to inconsistent conclusions in contexts which are ϵ-inconsistent (see section 2.4). For instance, a background encoding two defaults "birds fly" and "birds do not fly" gives rise to two contradictory conclusions $\mathtt{bird}(\mathtt{tim}) \mathrel{|\!\sim_K} \mathtt{fly}(\mathtt{tim})$ and $\mathtt{bird}(\mathtt{tim}) \mathrel{|\!\sim_K} \neg\mathtt{fly}(\mathtt{tim})$. This behavior does not arise in most default reasoning frameworks as these defaults

[8]This example is due to Matt Ginsberg.

are regarded as conflicting rather than inconsistent. The present framework makes such a distinction. This distinction is useful as, usually, inconsistent default theories reveal something wrong about the encoding. Three defaults such as $p \rightarrow q$, $p \wedge q \rightarrow r$ and $p \rightarrow \neg r$ constitute another example of inconsistent defaults.

The examples above illustrate some of the characteristics of the system of defeasible inference defined by rules 1–6. While the examples are treated satisfactorily, it should not be inferred that rules 1–6 provide an adequate formal account of default inference. They rather provide an account of some of the aspects of default reasoning, such as *specificity*, *cumulativity*, and the like, which belong to what we call the *conditional* dimension of defaults. There is another dimension to default reasoning, however, in which the present account has little to say. We call such a dimension the *causal* dimension of defaults, as it normally pops up in scenarios which involve causal relations. The Yale shooting scenario [Hanks and McDermott, 1986], for instance, belongs to such class. In chapter 5 we will analyze several such examples and construct a more refined account of defeasible inference in which both the *causal* and *conditional* aspects of defaults are considered.

Before proceeding with such an account, however, we will further investigate in the next two chapters the foundations of the system presented.

2.7 Related work

Rules 2–5 of the core are essentially equivalent to the logic of indicative conditionals developed by Adams [1966, 1975]. These rules also appear, in different forms, in most logics of conditionals (see [Nute, 1984]), where they are usually justified in terms of possible worlds rather than high probabilities. More recently, Makinson [1989] and Kraus *et al.* [1988] have worked a system that is equivalent to the core of **P**, but they derive on model-theoretic grounds. A common characteristics of all these proposals, though, is that they do not go beyond the core. So, while they display *non-monotonic* behavior in the evidence set, they remain *monotonic* in the set of defaults (conditionals); namely, the more defaults, the more inferences that are sanctioned (see theorem 2.2). However, the fact that the core can also be justified on model-theoretic grounds is by itself interesting, and as we will see in the next two chapters, fruitful.

In AI, most non-monotonic logics require the explicit addition of preferences in order to properly deal with interacting defaults [Reiter and Criscuolo, 1983].

In recent years, however, several novel systems of defeasible inference have been proposed which attempt to uncover such preferences.

In this regard, the closest formalism to **P** is perhaps the one proposed by Pollock [1988]. Pollock also combines probabilities and arguments, though in a different manner. He relates defeasible inference to what philosophers have called direct inference: the inference of *definite probabilities* relating members of certain classes, from *indefinite probabilities* relating the classes themselves. Pollock further regards a conditional indefinite probability $P(\mathrm{B}|\mathrm{A}) \geq \mathtt{r}$, for some $\mathtt{r}$ reasonably high, together with a given instance $\mathtt{a}$ of A, as constituting a *prima facie reason* (i.e. defeasible reason) for believing $\mathtt{a}$ to be a B. Prima facie reasons combine to form arguments, and undefeated arguments support what he calls warranted conclusions. The rest of his account is concerned with the conditions for argument defeat.

The main feature that distinguishes **P** from Pollock's framework, is the syntactic form taken by the former. Unlike Pollock's account, **P** constitutes a *calculus* of defeasible inference. Its simplicity is a result of both the focus on arbitrarily high and low probabilities, and the limited use of arguments for identifying independence assumptions. Additionally, Pollock's account relies on a non fully specified notion of 'projectibility,' as a result of not distinguishing 'primitive' from 'derived' defaults [Pollock, 1988].

A system close in form to the one proposed here is Delgrande's [1987]. Paralleling the correspondence between **P** and Adams' logic of indicative conditionals, Delgrande's system shares its core with a variant of the logics of counterfactuals. Delgrande's default logic is grounded on a possible world semantics rather than on probabilities. Still, such semantics does not circumvent the need for supplementing the system core with assumptions about independence. In this regard, while we characterize the notion of irrelevance in terms of arguments and embed it in the meta-predicate $I_K(\cdot)$, Delgrande appeals to fixed point constructions, used to generate new defaults and assertions which are added to the original set.

In Loui's [1987a] system, default reasoning emerges from dialectical argumentation. A set of rules are used to evaluate arguments in terms of syntactic attributes, like 'has more evidence', 'is more specific', etc. This set of rules appears to embed most of the inference rules that define our system and can be mostly justified in terms of them. Still, it is possible to find some differences. One such difference is that Loui's system is not (deductively) closed. It is possible to believe propositions A and B, and still fail to believe their conjunction [Loui, 1987a]. In our scheme, the deductive closure of believed propositions is established by theorem 1. Similarly,

Loui's preference for arguments based on 'more evidence' sometimes contradict our augmentation rule, as the confirmation of facts expected to hold might produce changes in belief.

Touretzky's [1986] account was motivated on the problems caused by redundant paths in inheritance networks. He was one of the first to suggest the use of specificity relations for filtering spurious ambiguities in default theories. Rather than a calculus of defeasible inference however, Touretzky's inferential distance can be regarded as a refinement of Reiter's default logic; to determine whether a proposition follows from a network, it is necessary to test whether the proposition holds in all the remaining extensions. Similar observations apply to Poole's [1985] specificity selection criterion.

Nute [1986], and Horty *et al.* [1987], on the other hand, define defeasible inference inductively, with special attention paid to 'specificity' relations. Horty *et al.* define a 'skeptical inheritance' scheme for homogeneous (defaults only) inheritance hierarchies, while Nute's system deals with linear arguments comprised of both defeasible and undefeasible rules. However, while Horty *et al.* rely on defaults to establish specificity relations, Nute relies only on 'strict' rules. In that regard, the difference between 'strict' rules and facts that Nute postulates, is reminiscent of the distinction made in [Poole, 1985] and [Delgrande, 1987] between necessary and contingent facts, and the one made here between background and evidence.

Chapter 3

High Probabilities and Preferential Structures

3.1 Introduction

In the previous chapter we described a system of defeasible inference made up of six rules. We showed that the five rules in the *core* can be given a probabilistic interpretation which guarantees that only highly probable conclusions can be derived from highly probable premises. The *irrelevance* rule, on the other hand, supplements the core with assumptions about independence. The resulting system captures a variety of patterns of default inference while providing a new vantage point from which default reasoning can be understood.

Interestingly, it is possible to justify the core of **P** on non-probabilistic grounds as well. In this chapter we present an alternative validation that rests on purely model-theoretic grounds, showing that ϵ-entailment is equivalent to a form of *preferential entailment*. Preferential entailment is a generalization of classical entailment in which the truth of the target conclusion is considered only over the *preferred* models of the premises, rather than over *all* their models [Shoham, 1988]. The framework of preferential entailment underlies the semantics of circumscription as proposed by McCarthy [1980, 1986], and the notion of subimplication due to Bossu and Siegel [1985]. It is also closely related to the possible world semantics of counterfactual logics, and to Lewis'[1973, section 2.3] comparative similarity formulation in particular.

The correspondence between ϵ-entailment and preferential entailment will pave the way for a better understanding of both the potentials and limitations of interpretations of defaults structured around the notions of high probabilities and preferred models. Additionally, it will furnish us with a completeness characterization of the inference rules that constitute the core, as well as with the building blocks for an extension of the core taken up in chapter 4.

This chapter largely reformulates results that go back to Adams [1966, 1975], and relies on recent developments in the area of preferential logics due to Kraus *et al.* [1988], Makinson [1989], and Lehmann and Magidor [1988], and extensions due to Pearl [1989b].

3.2 Preferential Structures and p-entailment

As discussed in section 1.5, the circumscription of a predicate P in a theory T is a second order formula that asserts that the tuples which can be shown to comply with P in T are the *only* tuples that do. Model-theoretically, the effect of circumscription is to exclude from consideration all those models of T which assign an extension to P larger than necessary. The models left assign a *minimal* extension to P, and thus the name *minimal models*. The formulas entailed by the circumscription of P in T are then simply the formulas which hold in the models of T minimal in P. It is also common to say that these formulas are *minimally entailed* by T, where the minimality criterion is understood relative to the extension of P.

Minimal entailment can be regarded as a generalization of classical logical entailment. While a proposition A *logically* entails a proposition B when B is true in all models of A, A *minimally* entails B when B is true in all models of A considered *minimal* in some sense. The notion of minimality underlying circumscription is a function of both the extension of the circumscribed predicates and the interpretation of fixed predicates and function symbols ([Lifschitz, 1985, Etherington, 1988]). However, other minimality criteria which do not necessarily translate into a simple circumscriptive axiom, are also possible. Shoham [1986], for instance, developed an alternative minimization criterion motivated by the apparent limitations of circumscription for handling problems in the temporal domain [Hanks and McDermott, 1986].

Shoham [1988] later investigated the properties associated with the form of

entailment that results from an abstract 'minimality' criterion; namely, a simple strict partial order on interpretations called the *preference* order. Shoham found that the resulting form of entailment, called preferential entailment, while being non-monotonic preserves yet certain traits of classical logic. Further properties of preferential entailment have been recently established by Kraus *et al.* [1988] and Makinson [1989]. In the recapitulation below, our notation and terminology is closest to Kraus *et al.*

We assume an underlying language $\mathcal{L}$ and a space $\mathcal{I}_\mathcal{L}$ of classical interpretations defined over $\mathcal{L}$. Default theories $T = \langle K, E \rangle$, as in chapter 2, are composed of a background context $K = \langle L, D \rangle$ and a set E of evidential sentences. A model of a default theory T is to be understood as an interpretation that satisfies the sentences in both L and E. The defaults in D will determine how such models should be ordered.

Definition 3.1 *A* preferential model structure *(p-structure) is a pair $\langle \mathcal{I}, < \rangle$, where $\mathcal{I}$ denotes a non-empty collection of interpretations, $\mathcal{I} \subseteq \mathcal{I}_\mathcal{L}$, and '$<$' denotes an irreflexive and transitive order relation over $\mathcal{I}$ called the* preference order.

Within a particular p-structure $\langle \mathcal{I}, < \rangle$, we usually read the the notation $M < M'$ for two interpretations M and M' in $\mathcal{I}$, as saying that M is *preferred* to M'. Furthermore, when M is a model of T and there is no model of T preferred to M in $\mathcal{I}$, we will say that M is a *preferred model* of T:

Definition 3.2 *A model M of a default theory T is a* preferred *model of T in a p-structure $\langle \mathcal{I}, < \rangle$, iff $M \in \mathcal{I}$ and there is no model M' of T in $\mathcal{I}$ such that $M' < M$.*

The semantics which underlies circumscription, for instance, can be understood in terms of a particular type of preferential model structure $\langle \mathcal{I}_{\text{circ}}, <_{\text{circ}} \rangle$, in which $\mathcal{I}_{\text{circ}}$ corresponds to the set of all logically possible interpretations, and the ordering '$<_{\text{circ}}$' on interpretations is such that an interpretation M is preferred to an interpretation M' when both M and M' coincide on the domains and the interpretation of fixed predicate and function symbols, but M yields a smaller extension for the circumscribed predicates [Lifschitz, 1985, Etherington, 1988].

In a way analogous to circumscription, preferential entailment could be defined in terms of the truths in the *preferred* models of a given theory. The problem,

however, is that the existence of preferred models for consistent theories is not always guaranteed. There may be p-structures which contain models of T but no *preferred* models of T, and therefore, which would entail any sentence in the language. This has been noted in the case of circumscription by Etherington *et al.* [1985] who show that in certain circumstances, the circumscription of a consistent theory may turn out to be inconsistent. They also identify a class of *well-founded* theories in which such a behavior is guaranteed not to arise. This well-foundedness condition has an immediate generalization for arbitrary preferential structures which we express as follows:

Definition 3.3 *A default theory T is* well-founded *relative to a given preferential model structure $\langle \mathcal{I}, < \rangle$, iff for every model M of T in $\mathcal{I}$ there is a* preferred *model M' of T in $\mathcal{I}$ such that $M' < M$ or $M' = M$.*

In other words, a theory is well-founded in a given structure when for every *non-preferred* model M, there is a *preferred* model M' preferred to M. In certain cases, this well-foundedness condition can be tested syntactically. Bossu and Siegel [1985], for instance, showed that *universal* (classical) theories are well-founded in certain circumscriptive preferential structures. Here, and very much like Kraus *et al.* [1988] and Makinson [1989], we will focus on the study of a special kind of preferential model structures for which *every* possible default theory is well-founded. We call these preferential model structures, *well-founded p-structures*:[1]

Definition 3.4 *A preferential model structure $\pi = \langle \mathcal{I}, < \rangle$ is* well-founded *relative to a* background context *K when every theory of the form $T = \langle K, E \rangle$ is well-founded relative to π.*

Preferential structures $\langle \mathcal{I}, < \rangle$ which do not involve infinite descending chains; namely, interpretations M_i, such that $M_{i+1} < M_i$ for every positive i, will be thus well-founded. In particular, preferential model structures defined over finite propositional languages $\mathcal{L}$ will be well-founded. Still the condition of well-foundness are very strong, and will be later relaxed in chapter 4.

The importance of well-founded structures is twofold. First, every logically consistent default theory $T = \langle K, E \rangle$ will have a non-empty set of preferred models in every well-founded preferential model structure. Thus, an entailment relation defined in terms of the preferred models of T will be guaranteed to be consistent

[1] Analogous structures are called 'smooth' by Kraus *et al.*, and 'stoppered' by Makinson.

as long as T itself is logically consistent. Furthermore, as the following theorem states, any such entailment relation will obey many of the rules of the probabilistic system **P** studied in chapter 2.

Theorem 3.1 (Kraus et al.) *If the expression $E \mathrel{|\!\sim_K} p$ is interpreted as asserting that p is true in all the preferred models of $T = \langle K, E\rangle$ in every preferential model structures well-founded relative to K, then the following rules are sound:*

Rule 2 (Deduction) If $E \models_K p$ then $E \mathrel{|\!\sim_K} p$

Rule 3 (Augmentation) If $E \mathrel{|\!\sim_K} p$ and $E \mathrel{|\!\sim_K} q$ then $E, p \mathrel{|\!\sim_K} q$

Rule 4 (Reduction) If $E \mathrel{|\!\sim_K} p$ and $E, p \mathrel{|\!\sim_K} q$ then $E \mathrel{|\!\sim_K} q$

Rule 5 (Disjunction) If $E, p \mathrel{|\!\sim_K} r$ and $E, q \mathrel{|\!\sim_K} r$ then $E, p \vee q \mathrel{|\!\sim_K} r$

In other words, well-founded preferential model structures provide an interpretation alternative to ϵ-entailment under which rules 2–5 of **P** are valid. Note, however, that absent from these rules is rule 1. Rule 1 is a critical rule in the core as it is the only one which takes into account the default component of the context in question. The failure of rule 1 to hold, however, is not surprising. We have only talked so far about models of theories $T = \langle K, E\rangle$, in which the default component D of $K = \langle L, E\rangle$ plays no role. The role of D indeed will not be in determining what the models of T are, but in determining how such models are supposed to be ordered. Preferential model structures which comply with such order will be said to be *admissible:*

Definition 3.5 *A well-founded preferential model structure $\langle \mathcal{I}, <\rangle$ relative to a background context $K = \langle L, D\rangle$ is* admissible *with K iff every interpretation in $\mathcal{I}$ satisfies L, and for every default $p \rightarrow q$ in D, (a) q is true in all preferred models of p in $\mathcal{I}$, and (b) there is an interpretation in $\mathcal{I}$ that satisfies p.*

Preferential entailment is defined in terms of the preferred models of a body of evidence within the preferential structures *admissible* with its background context.

Definition 3.6 *A default theory $T = \langle K, E\rangle$* preferentially entails (p-entails) *a sentence p iff p is true in all the preferred models of E in every preferential model structure admissible with K.*

Note that within any preferential model structure admissible with K, a model of E will also be a model of $T = \langle K, E \rangle$, and vice versa. We emphasize that the preferred interpretation must be models of E rather than of $T = \langle K, E \rangle$, to highlight the similarity between *preferential entailment* and *ϵ-entailment* (chapter 2): while in ϵ-entailment the background K picks the admissible probability distributions which are then conditioned upon the evidence E, in p-entailment, the background K picks the admissible preferential model structures, from which the preferred models of E are selected.

The restriction to *admissible* p-structures yields a model-theoretic entailment relation which provides an alternative validation of the core of **P**:

Theorem 3.2 (Soundness) *If the proposition p is is derivable from a context $T = \langle K, E \rangle$ by means of rules 1–5, then p is preferentially entailed by $T = \langle K, E \rangle$.*

The simple nature of rules 1–5, together with the existence of natural justifications in terms of both probabilities and models, have lead some researchers to present the core as a minimal default inference shell (e.g. [Pearl, 1989a]). We will return to this theme in section 3.7. Meanwhile we will analyze in more detail what are the features that make a probabilistic interpretation that relies on high conditional probabilistic statements, and a model-theoretic interpretation that relies on a preference relation on models, legitimize a common set of inferences. In particular, we would like to know whether this set of inferences is *complete* with respect to either ϵ-entailment or p-entailment, and whether this latter two entailment relations are indeed equivalent. These two topics will constitute the subject of the remainder of this chapter. We will show first that ϵ-entailment and p-entailment are indeed *equivalent*, and then, that the core is not only sound with respect to them, but also *complete*. In different ways, similar results have been shown by Adams [1966, 1975], Kraus *et al.* [1988], Makinson [1989], and Lehmann and Magidor [1988].

We will show the equivalence between ϵ-entailment and p-entailment, by exploiting the relation between entailment and consistency established before for ϵ-entailment. *P-consistency* is defined in a way analogous to ϵ-consistency:

Definition 3.7 *A background context K is* p-consistent *iff there is a preferential model structure admissible with K. Otherwise, K p-inconsistent.*

For example, an ϵ-inconsistent background context K containing two defaults $p \rightarrow q$ and $p \rightarrow \neg q$ is p-inconsistent. A preferential model structure admissible

with K must accommodate a non-empty set of preferred models of p which must satisfy q, due to the presence of the default $p \rightarrow q$, and $\neg q$, due to the presence of the default $p \rightarrow \neg q$. Hence, there cannot be p-structures admissible with K, and thus, K is p-inconsistent.

This example suggests that the conditions for ϵ-consistency and p-consistency may be closely related. We will investigate such a relation in detail in section 3.4. Some preliminary results, however, will be needed. First, note that the relation between ϵ-entailment and ϵ-consistency established in lemma 2.4, also holds for preferential structures:

Lemma 3.1 *A default theory $T = (K, \{p\})$ with a background context $K = (L, D)$ p-entails a sentence q if only if the background $K' = (L, D \cup \{p \rightarrow \neg q\})$ is p-inconsistent.*

3.3 Layered Structures and l-entailment

To get a deeper insight into the relation between p-entailment and ϵ-entailment we need to relate the structures on which such notions rely. The preferential model structures deal with full fledged *interpretations* which are *partially ordered*; probability distributions, on the other hand, deal with *worlds* ordered by their probability ranks.[2] We will thus find useful to introduce an intermediate class of structures, called *layered world structures*, consisting of non-empty subsets of $\mathcal{W}_\mathcal{L}$, the set of all possible worlds, ordered according to ranking function (see also [Lehmann and Magidor, 1988]).

Definition 3.8 *A* layered world structure *(l-structure) is a pair $(\mathcal{W}, \kappa)$, where $\mathcal{W}$ is a non-empty set of worlds, $\mathcal{W} \subseteq \mathcal{W}_\mathcal{L}$, and κ is a function which assigns a non-negative integer to each world in $\mathcal{W}$.*

Layered world structures are structures which comprise a set of worlds $\mathcal{W}$ organized in layers $\mathcal{W}_0, \mathcal{W}_1, \ldots, \mathcal{W}_i, \ldots$. Every world W in $\mathcal{W}$ belongs to a single layer $\mathcal{W}_i$, whose index i represents the rank $\kappa(W)$ of W.

[2] A world is a truth valuation over the sentences of the language. Every interpretation is associated with a world, though, in first order languages, a single interpretation will usually be associated with many worlds.

The definition of preferred worlds within a given l-structure is analogous to the definition of preferred models within a given p-structure:

Definition 3.9 *A world W that satisfies a default theory T is a* preferred *world of T in a l-structure $\langle \mathcal{W}, \kappa \rangle$ iff $W \in \mathcal{W}$ and there is no world W' in $\mathcal{W}$ that satisfies T for which $\kappa(W') < \kappa(W)$.*

We follow the aforementioned convention of saying that a world W satisfies a default theory $T = \langle K, E \rangle$ with a background $K = \langle L, D \rangle$, when W satisfies the formulas in both L and E.

Note that the ordering of worlds in terms of non-negative integer rankings ensures the existence of a *minimum* ranked set of worlds among the worlds satisfying any logically consistent theory. As a result, layered world structures, unlike preferential model structures, are always well-founded.[3]

The admissibility of layered world structures and the notions of l-entailment and l-consistency are defined in a way analogous to p-structures:

Definition 3.10 *A layered world structure $\langle \mathcal{W}, \kappa \rangle$ is* admissible *with a background $K = \langle L, D \rangle$ iff every world in $\mathcal{W}$ satisfies L, and for every default $p \rightarrow q$ in D, (a) q is true in all preferred worlds of p in $\mathcal{W}$, and (b) there is a world in $\mathcal{W}$ that satisfies p.*

Definition 3.11 *A default theory $T = \langle K, E \rangle$* l-entails *a sentence p iff p is true in all the preferred worlds of T of every layered world structure admissible with K.*

Definition 3.12 *A background K is* l-consistent *iff there is a layered world structure admissible with K. Otherwise, K is l-inconsistent.*

The relation between entailment and consistency found for ϵ-semantics and p-entailment also holds for l-entailment:

Lemma 3.2 *A default theory $T = \langle K, \{p\} \rangle$ with a background $K = \langle L, D \rangle$ l-entails a sentence q if only if $K' = \langle L, D \cup \{p \rightarrow \neg q\} \rangle$ is l-inconsistent.*

[3]This is one of the few differences between our layered world structures and Lehmann's and Magidor's [1988] ranked models. Notions similar to layered world structures also appear under the name of "P-orderings" in Adams [1966], and more recently, in Spohn's [1988] conditional functions.

3.4 Equivalences

Admissible layered world structures will play the role of a bridge between admissible preferential model structures and admissible probability distributions. On the one hand, admissible layered world structures are closely related to admissible preferential model structures due to the relation between worlds and interpretations, and rankings and partial orders; on the other, admissible layered world structures are closely related to admissible probability distributions due to the relation between world ranks and world probabilities. The results below make such correspondences precise for *finite propositional languages*.

Lemma 3.3 *A background K is p-consistent if and only if K is l-consistent.*

Lemma 3.4 *A background K is ϵ-consistent if and only if K is l-consistent.*

In light of the results relating entailment and consistency across preferential model structures, layered world structures and probability distributions, these correspondences will thus permit us to express the relations between ϵ-entailment, l-entailment and p-entailment in the following form:

Theorem 3.3 (Equivalences) *Let $K = \langle L, D \rangle$, and $K' = \langle L, D + \{p \rightarrow \neg q\} \rangle$ be two background contexts, and let $T = \langle K, \{p\} \rangle$ be a default theory. Then, for finite propositional languages, the following statements are equivalent:*

(1) T ϵ-entails q
(2) K' is ϵ-inconsistent
(3) K' is l-inconsistent
(4) T l-entails q
(5) K' is p-inconsistent
(6) T p-entails q

Figure 3.1 illustrates the resulting relations. It is clear now *why* the core of **P**, justified originally in terms of probabilities, is also valid under a preferential interpretation: there is a two way correspondence between the structures that underlie both semantic accounts, and whenever one structure renders a given inference invalid in one interpretation, a corresponding structure can be constructed which renders the same inference invalid in the other interpretation.

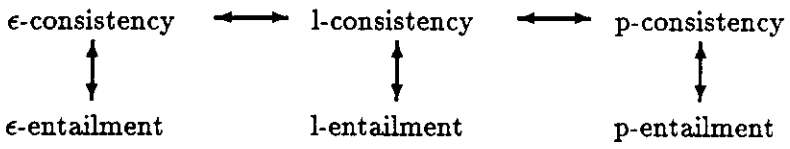


Figure 3.1: Equivalence between various forms of entailment

In particular we can now prove the core of **P** to be *complete* with respect to ϵ-entailment or p-entailment, by proving it to be complete with respect to l-entailment. That is indeed what we are going to do in the rest of this chapter. First, however, we will introduce the notion of *default rankings*.

3.5 Default Rankings

A *default ranking* over a background context $K = \langle L, D \rangle$ is a function that assigns a non-negative integer to each default in D. Intuitively, we may think the rank of a default as a measure of its strength. In this regard, default rankings are comparable to priorities in McCarthy's [1986] prioritized circumscription. However, rather as a means for the user to *express* the strength of defaults, we are going to use default rankings to *uncover* them from the information in the background context in which they appear. For instance, in a consistent background containing two defaults $\mathtt{p} \rightarrow \mathtt{q}$ and $\mathtt{p} \wedge \mathtt{r} \rightarrow \neg\mathtt{q}$, the more 'specific' default $\mathtt{p} \wedge \mathtt{r} \rightarrow \neg\mathtt{q}$ will automatically receive a higher rank than the less 'specific' default $\mathtt{p} \rightarrow \mathtt{q}$. Such a use of default rankings was implicit in the original work of Adams [1966], and has been recently exploited by Pearl [1989b] in an extension of Adams' work.

We will refer to default rankings that comply with the preferences implicit in K as *admissible default rankings*. To make such preferences precise, let us say that a world W *verifies* a default $p \rightarrow q$ in K if W satisfies both p and q, and that W *falsifies* $p \rightarrow q$, if it satisfies p but fails to satisfy q [Adams, 1975]. Furthermore, let us define the notion of conflict among defaults as follows:[4]

[4]The conditions under which *a set of defaults* $p_i \rightarrow q_i$ is in conflict with a default $p \rightarrow q$ are closely related to the conditions under which *a set of assumptions* is an conflict with a default (section 2.5), provided that the assumptions are drawn from the material counterparts $p_i \Rightarrow q_i$ of the defaults $p_i \rightarrow q_i$.

Definition 3.13 *Let $p \rightarrow q$ and $p_i \rightarrow q_i$, $i = 1, \ldots, n$, be a set of defaults in a background context K. Then the default $p \rightarrow q$ is said to be in* conflict *with* the set of defaults $p_i \rightarrow q_i$, $i = 1, \ldots, n$ *iff* $p \nvdash_K \neg(p \Rightarrow q) \vee \neg(p_1 \Rightarrow q_1) \vee \cdots \vee \neg(p_n \Rightarrow q_n)$.

Namely, a default $p \rightarrow q$ is in conflict with a set D' of defaults when the verification of $p \rightarrow q$ in K amounts to the falsification of some default in D'. Moreover, since in the preferred worlds of p in any structure admissible with K the default $p \rightarrow q$ must be verified, it is reasonable to assume that in any such structure it is preferable to falsify one of the defaults in D' than $p \rightarrow q$. Admissible default rankings are defined to reflect such preferences:[5]

Definition 3.14 *Let σ denote a bounded default ranking over a background context $K = \langle L, D \rangle$, and let $\sigma(D')$, for a subset D' of D, stand for the rank of the minimally ranked default in D' if D' is non-empty, and for infinite otherwise. Then σ is a default ranking* admissible *with K, iff for every default $p \rightarrow q$ in D and every set D' of defaults in conflict with $p \rightarrow q$ in K, $\sigma(D') < \sigma(p \rightarrow q)$ holds.*

For instance, if K contains a default $\mathtt{p} \rightarrow \mathtt{q}$ and a second default $\mathtt{p} \wedge \mathtt{r} \rightarrow \neg\mathtt{q}$ in conflict with the former, a default ranking σ will admissible with K, only if $\sigma(\mathtt{p} \wedge \mathtt{q} \rightarrow \mathtt{r}) > \sigma(\mathtt{p} \rightarrow \neg\mathtt{r})$. More generally, whenever $\mathtt{p} \rightarrow \mathtt{q}$ and $\mathtt{p}' \rightarrow \neg\mathtt{q}$ are two defaults in K, such that p is 'more specific' than p', i.e. $\mathtt{p} \nvdash_K \mathtt{p}'$, then the default $\mathtt{p} \rightarrow \mathtt{q}$ will have a higher rank than $\mathtt{p}' \rightarrow \neg$ in every ranking admissible with K.

Note that it is simple to come up with background contexts which do not accept *any* admissible default ranking. For instance, no default ranking will be admissible with K if K contains a pair of defaults $p \rightarrow q$ and $p \rightarrow \neg q$. Indeed, the existence of admissible rankings turns out to be a sufficient and necessary condition for the existence of admissible structures:

Theorem 3.4 *A background context $K = \langle L, D \rangle$ is consistent if and only if there is a default ranking admissible with K.*

In light of the correspondences between entailment and consistency established in the previous section, default rankings thus become an alternative way to evaluate entailment. Default rankings will be particularly convenient for such task due to

[5] A default ranking σ over a background K is *bounded*, when there is a constant k, such that for every default $p \rightarrow q$ in K, $\sigma(p \rightarrow q) \leq k$.

the explicit relation between the admissibility conditions and the syntactic form of K.

To show these advantages we will introduce yet another syntactic notion; the notion of *default clashes.*

Definition 3.15 *A non-empty set of defaults D', $D' \subseteq D$, constitutes a* clash *in a background $K = \langle L, D \rangle$, iff every default $p \rightarrow q$ in D' is in conflict with D' in K. Likewise, a default $p \rightarrow q$* clashes *with D', if the set $D' + \{p \rightarrow q\}$ constitutes a* clash *in K.*

Two defaults $p \rightarrow q$ and $p \rightarrow \neg q$, for instance, clash in any background context. Indeed, clashes of defaults are the only reason for inconsistency. In other words, we can test the consistency of a background context by purely syntactic means by testing the presence of default clashes:[6]

Lemma 3.5 *A background context is consistent if and only if it does not contain a clash.*

Furthermore, since we can evaluate whether a proposition q is entailed by a proposition p in a background context $K = \langle L, D \rangle$ by testing the consistency of the background context $K' = \langle L, D + \{p \rightarrow \neg q\} \rangle$, clashes provide us with the facility to test *entailment* by syntactic means (fig. 3.2):

Lemma 3.6 *In a background context $K = \langle L, D \rangle$ p entails q if and only if the background $K' = \langle L, D + \{p \rightarrow \neg q\} \rangle$ contains a clash.*

Moreover, it is possible to test whether a background context $K = \langle L, D \rangle$ contains a clash in a 'greedy' fashion. Namely, if D itself is a clash then every default $p \rightarrow q$ in D must be in conflict with D; otherwise, every default *not* in conflict with D is guaranteed not to participate in any clash and can thus be removed, leaving a smaller set D' which can be tested by similar means. As originally noted by Pearl [1989b], such a procedure, together with the result summarized in lemma 3.6, permits computing entailment in time polynomial in the number of defaults in D, provided unit time satisfiability tests:

[6] Lemma 3.5 assumes that the background context K contains a finite number of default schemas $p(x) \rightarrow q(x)$ such that any two instances gives rise to identical conflicts except for term substitutions. This permits to assign the same rank to all the potentially infinite instances of a given default schema.

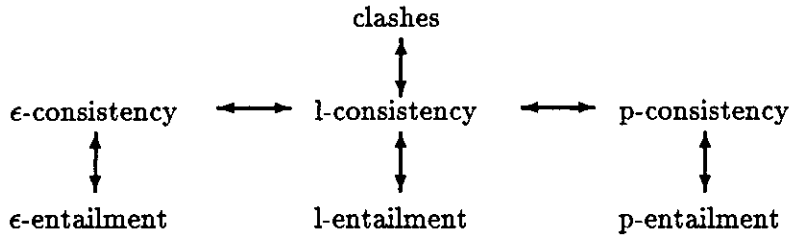


Figure 3.2: Entailment and default clashes

Theorem 3.5 (Pearl) *For a background context $K = \langle L, D \rangle$ with n defaults, there is a $\mathcal{O}(C(n) \times n^2)$ procedure for testing whether a sentence q is entailed by a sentence p in K, where $C(n)$ is the complexity associated with testing the satisfiability of n sentences in the language fragment that contains the sentences in L, the material counterparts of the defaults in D, and the material counterpart $p \Rightarrow \neg q$ of the default $p \rightarrow \neg q$ (e.g. $C(n) = \mathcal{O}(n)$ for Horn clauses).*

3.6 Completeness results

Provided with the results about default rankings and clashes, we are now ready to attack that last problem of this chapter: the completeness of the core relative to ϵ-entailment and p-entailment. We will however introduce a final notion which will help us simplify matters considerably. This is the notion of *quasi-conjunctions* originally introduced in [Adams, 1975].

The quasi-conjunction $C(D)$ of a set D of defaults $p_i \rightarrow q_i$, $i = 1, \ldots, n$, is the default: $p_1 \vee p_2 \vee \cdots \vee p_n \rightarrow (p_1 \Rightarrow q_1) \wedge (p_2 \Rightarrow q_2) \wedge \cdots \wedge (p_n \Rightarrow q_n)$. As noted by Adams, quasi-conjunctions permit us to map the problem of whether a sentence q is entailed by a sentence p in a background context $K = \langle L, D \rangle$, by the problem of testing the consistency of a simpler background context $K' = \langle L, D' \rangle$ containing only two defaults: the quasi-conjunction $C(D)$ of D, and the denial $p \rightarrow \neg q$ of the default $p \rightarrow q$. This feature is a consequence of results established in the previous section and the following lemma:

Lemma 3.7 *Let $p \rightarrow q$ be a default in D, let D' a subset of D, and let $C(D')$ be the quasi-conjunction of D'. Then, $p \rightarrow q$ clashes with D' in a background context*

$K = \langle L, D\rangle$, if and only if $p \rightarrow q$ clashes with $C(D')$ in the background context $K' = \langle L, D''\rangle$ with $D'' = \{C(D'), p \rightarrow q\}$.

The lemma is a simple consequence of the logical equivalence between the sentence

$$\neg(p_1 \Rightarrow q_1) \vee \cdots \vee \neg(p_n \Rightarrow q_n)$$

and the sentence

$$\neg(p_1 \vee p_2 \vee \cdots \vee p_n \Rightarrow (p_1 \Rightarrow q_1) \wedge (p_2 \Rightarrow q_2) \wedge \cdots \wedge (p_n \Rightarrow q_n))$$

With this final result, we are now ready to prove that the core of **P** as embodied in the rules 1–5 is *complete* with respect to all three forms entailment considered for *finite propositional languages*. We use the notation '$p \mathrel{\overset{\circ}{\sim}_K} q$' to express that q is derivable from p in K by means of rules 1–5.

Theorem 3.6 (Completeness) *If p entails q in a* consistent *background $K = \langle L, D\rangle$, then $p \mathrel{\overset{\circ}{\sim}_K} q$.*

Proof Note first, that if p entails q in K, the background context K' that results from the addition of $p \rightarrow \neg q$ to K must be inconsistent. Furthermore, since K is assumed to be consistent, the results above imply that the default $p \rightarrow \neg q$ must clash with a subset D' of D in L, and therefore, that $p \rightarrow \neg q$ must clash with the quasi-conjunction $C(D')$ of D'. The rest of the proof is a straightforward consequence of the following two results:

Lemma 3.8 *Let $K = \langle L, D\rangle$ be a background context, and D' be a non-empty subset of D. Then, if $r \rightarrow s$ stands for the quasi-conjunction $C(D')$ of D', $r \mathrel{\overset{\circ}{\sim}_K} s$.*

Lemma 3.9 *Let $K = \langle L, D\rangle$, and $K' = \langle L, D'\rangle$ be two background contexts sharing the same set L of sentences. If $p \rightarrow \neg q$ clashes with $r \rightarrow s$ in K' and $r \mathrel{\overset{\circ}{\sim}_K} s$, then $p \mathrel{\overset{\circ}{\sim}_K} q$.*

Thus, we have that the core is complete with respect to entailment in *consistent* background contexts. The consistency condition is required because the semantic accounts legitimize any sentence in the language when K is inconsistent, while the core does so only in certain contexts. For example, from an inconsistent background context K containing two defaults $p \rightarrow q$ and $p \rightarrow \neg q$ the core will not derive $E \mathrel{|\!\sim^{\circ}_{K}} \mathbf{false}$, unless it can previously derive $E \mathrel{|\!\sim^{\circ}_{K}} p$.

Actually, there are various ways in which the consistency requirement can be dropped from the completeness theorem. One way is to legitimize arbitrary derivations in inconsistent contexts. This can be accomplished by means of an additional rule of inference which permits the special atom **false** to be derived when an inconsistency in K is detected [Adams, 1975]. A second, more appealing option consist of relaxing the admissibility requirements on structures and probability distributions. This approach has been pursued in [Kraus *et al.*, 1988], and amounts to dropping the requirement that every default be verified in some world. In terms of probabilities, this amounts taking into considerations non-proper probability distributions, for which it is agreed to set $P_K(q \mid p)$ equal to one when $P_K(p)$ is equal to zero (see [Adams, 1966]).

We have chosen the more stringent admissibility conditions together with the consistency properties that they entail, since they provide a simpler and more insightful correspondence between the probabilistic and model-theoretic accounts of defaults. Furthermore, the resulting consistency conditions impose a reasonable integrity constraint on defaults: pairs of defaults such as "birds fly" and "birds do not fly, " for instance, are ruled out as inconsistent.[7] Last but not least, these consistency conditions, expressed in terms of default rankings, will turn out to be essential for a semantic and proof-theoretic extension of the core to be developed in the next chapter.

3.7 Related Work

The soundness and completeness of rules 1–5 with respect to ϵ-entailment, were informally sketched in Geffner and Pearl [1987], and can be traced back to Adams [1966, 1975]. Preferential model structures and layered world structures, correspond, in essence, to structures recently advanced by Kraus *et al.* [1988], Makinson [1989] and Lehmann and Magidor [1988]. Lehmann and Magidor also noted the

[7] Even more stringent conditions on default and strict rules have been recently advanced in [Goldszmidt and Pearl, 1989].

connection between accounts based on infinitesimal probabilities and preferences among models; a connection which Adams himself explored in [Adams, 1978]. The notion of default rankings is due to Pearl [1989b]. As far as I know, this chapter is the first coherent and self-contained treatment of all these ideas.

A few remarks follow about the status of the inference rules legitimized by these accounts. From our discussion in section 2.5, it is clear that this set of inference rules, as well as the semantic accounts which render them sound and complete, do not provide a *complete* characterization of default reasoning. They fail to sanction inference patterns which involve independence assumptions as well as many other patterns to be studied in the next two chapters. What is not so clear, however, is whether these rules should be regarded as a *minimal* set of rules to be satisfied by any reasonable account of defeasible inference. In this regard, *augmentation* and *reduction*, also known as *cumulative monotony* and *cumulative transitivity* [Makinson, 1989], have attracted most of the attention. This pair of rules establish that two contexts $T = \langle K, E \rangle$ and $T' = \langle K, E + \{p\} \rangle$ legitimize the same conclusions, when p is a consequence of T. Early in 1985, Gabbay [1985] argued, on proof-theoretic grounds, that these rules, together with a weak form *deduction*, define minimal requirements on any reasonable non-monotonic consequence relation. Such a position has been lately echoed, on a semantic basis, by Kraus *et al.* [1988] and Pearl [Pearl, 1989a]. Our position is that while these rules are reasonable, they are not necessarily 'inescapable.' Indeed, both the probabilistic and model-theoretic accounts which validate cumulativity, embed questionable assumptions. The probabilistic interpretation, for instance, regards defaults as having arbitrarily high conditional probabilities. The model-theoretic account, on the other hand, regards the preference relation on models to be an exclusive function of the background context. Still, examples can be constructed whose intended behavior demands preferences to depend on *both* background and evidence (e.g. example 5.1) . However, once the space of admissible preferential structures is so determined, the cumulative behavior is no longer guaranteed.

Similarly, Lehmann and Magidor [1988] discuss a rule suggested by Makinson called *rational monotony*, which holds in what we have called layered world structures, but does not hold in preferential model structures. Rational monotony is a strong form of augmentation, which permits to carry a conclusion q from a context $T = \langle K, E \rangle$ to a context $T' = \langle K, E + \{p\} \rangle$, as long as the *negation* of p is *not* a consequence of T. An extension of preferential entailment based on rational monotony has been recently advanced in [Lehmann, 1989], on the grounds that reasonable accounts of defeasible inference should be as monotonic as possible. Whether this intuition is right, however, remains an empirical matter. For exam-

ple, a consequence relation obeying rational monotony will force us to conclude $\neg p'$ from p, and $\neg p$ from p', from any unresolved conflicting pair of defaults $p \rightarrow q$ and $p' \rightarrow \neg q$. Such a behavior, however, may turn out to be too adventurous, as when p is connected to p' via a 'diamond' structure (e.g. $p \rightarrow r \rightarrow p'$ and $p \rightarrow s \rightarrow \neg p'$). In such a case, though there are no grounds to conclude q over $\neg q$ given p and p', there are no grounds to conclude $\neg p'$ from p either. Still, a consequence relation obeying cumulativity and rational monotony is forced to make such a choice.

Chapter 4

Beyond High Probabilities and Preferential Structures

4.1 Defaults and Conditionals

The last two chapters have analyzed in detail two semantic accounts of defaults, one which relies on probabilities, the other which relies on a preference relation on models. Both interpretations legitimize a set of inferences identical to those sanctioned by the core of the system introduced in chapter 2. These inferences follow from regarding defaults as *conditional* assertions. While the probabilistic interpretation regards a default $p \rightarrow q$ as asserting that the probability of q, given that p *represents all the available evidence*, is high; the preferential interpretation regards $p \rightarrow q$ as asserting that q is true in *all the preferred models of* p. In both cases something is asserted about a particular context, constraining other contexts as well as a result of the axioms of probability theory in the first case, and as a result of the postulates on the preference relation on models in the second. The core is the logic of such constraints.

With a few exceptions (e.g., [Delgrande, 1987]), most work in non-monotonic logics has neglected this conditional dimension of defaults. Such a neglect has translated into important limitations, though also, into some important assets. The limitations arise from having to account for preferences among defaults by some other means. Thus many special ways of resolving counter-intuitive ambiguities have been proposed, ranging from those which rely on "specificity" considerations (e.g., [Poole, 1985, Nute, 1986, Loui, 1987a]) to those which rely on the user (e.g.,

prioritized circumscription [McCarthy, 1986, Lifschitz, 1988a], non-normal defaults [Etherington and Reiter, 1983]).

The benefits from ignoring the conditional dimension of defaults shows in the fact that non-monotonic logics (e.g. [Reiter, 1980, McDermott and Doyle, 1980, McCarthy, 1980, Moore, 1985b]) are able to capture intuitive default inferences which conditional interpretations of defaults cannot. In non-monotonic logics a default $p \rightarrow q$ is not viewed as an assertion q, bound to a particular context p,[1] but as a *prima facie reason* to assert q in *all* those contexts in which p holds. So, while these formalisms do not guarantee that q will follow when p is the only evidence available, because there may also be reasons for $\neg q$ in that context, they nevertheless induce a preference for q in *all* contexts in which p is true. This disposition escapes conditional interpretations of defaults where the success to account for specificity preferences is rooted in a clear-cut distinction between the context p, where a default $p \rightarrow q$ *guarantees* the truth of q, from contexts in which p holds, which are almost unconstrained by the presence of such a default.

In section 2.5 we discussed these limitations and developed an extension of the core around the notion of *irrelevance*. The idea was to strengthen the inferential import of defaults of the form $p \rightarrow q$ by guaranteeing the truth of q *both* in the context p, and in any context in which p holds and which is not suspected to contain evidence relevant to the negation of q. Such appeal to irrelevance considerations was an attempt to close the gap between the conditional and the traditional non-monotonic readings of defaults. Unfortunately, however, the irrelevance account presented in section 2.5 raises as many question as it solves. First of all it is not clear under what conditions the 'closure' of the core under the irrelevance rule is consistent. Even if so, we would like to know whether the syntactic account presented is adequate, and whether it can be justified on independent grounds, just as the core can be justified in terms of high probabilities and preferential structures.

We address these issues in this chapter by developing a form of entailment akin to preferential entailment, but which validates both the core and the irrelevance account. The new entailment relation is called *conditional* entailment, as it merges the conditional and traditional readings of defaults. The resulting interpretation takes us a step closer in our goal of uncovering the intended meaning and use of defaults, and constitutes the basis of further refinements to be studied in chapter 5.

[1] A context p refers here to a context in which p represents all the available evidence.

4.2 Closing the Gap: Conditional Entailment

In the previous chapter we have shown the core to be a sound and complete inference system with respect to ϵ-entailment and preferential entailment. In order to validate an extension to the core, it is necessary to weaken the conditions that define both entailment relationships. A natural choice is to restrict the space of admissible preferential structures or probability distributions in which valid conclusions must hold. In this chapter indeed we will study the patterns of inference sanctioned by a subset of admissible preferential structures. These structures will correspond to those admissible preferential model structures that can be induced from a given *default prioritization*. We will call the structures induced in this way *prioritized preferential structures*. Restricting our focus to default theories cast in *assumption based format*[2] will further permit us to replace reference to *default* priorities by the more standard notion of *assumption* priorities.

4.2.1 Model Theory

Prioritized preferential structures represent preferential model structures in which the relation that orders interpretations is determined by a given *priority order on assumptions*. Formally, if $\Delta[M]$ stands for the assumptions which are *false* under an interpretation M, *prioritized preferential structures* are defined as follows:

Definition 4.1 *A* prioritized preferential structure *is a quadruple* $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$, *where* $\mathcal{I}_\mathcal{L}$ *stands for the set of interpretations over the underlying language* $\mathcal{L}$, $\Delta_\mathcal{L}$ *stands for the set of assumptions in* $\mathcal{L}$, '$\prec$' *stands for an irreflexive and transitive priority relation over* $\Delta_\mathcal{L}$, *and* '$<$' *is a binary relation over* $\mathcal{I}_\mathcal{L}$ *such that for two interpretations* M *and* M', $M < M'$ *holds iff* $\Delta[M] \neq \Delta[M']$ *and for every assumption* δ *in* $\Delta[M] - \Delta[M']$ *there exists an assumption* δ' *in* $\Delta[M'] - \Delta[M]$ *such that* $\delta \prec \delta'$.

We will further assume that priority orderings '$\prec$' do not contain infinite chains; namely, there is no infinite sequence of assumptions $\delta_1, \delta_2, \ldots, \delta_i$ for which the relation $\delta_{i+1} \prec \delta_i$ holds for every positive i.

[2]Assumption based default theories are theories in which all defaults are of the form $p \rightarrow \delta$ for unique assumptions δ. See section 2.2 for details.

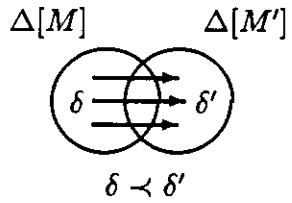


Figure 4.1: Ordering among interpretations in prioritized structures

The model structures defined in this way are called *preferential* because the priority relation '$\prec$' over $\Delta_{\mathcal{L}}$ induces a preferential model structure over the set $\mathcal{I}_{\mathcal{L}}$ of interpretations.

Lemma 4.1 *The quadruple $\langle \mathcal{I}_{\mathcal{L}}, <, \Delta_{\mathcal{L}}, \prec \rangle$ is a prioritized preferential structure only if the pair $\langle \mathcal{I}_{\mathcal{L}}, < \rangle$ is a preferential model structure.*

The ordering '$<$' on interpretations induced from the priority ordering '$\prec$' on assumptions regards the relation $\delta \prec \delta'$ as a preference to sustain the assumption δ' over the assumption δ in cases of conflict. A similar mapping from *predicate* priorities to ordering relations on interpretations occurs in Przymusinski's characterization of the perfect model semantics of general logic programs [Przymusinski, 1987] and in McCarthy's prioritized circumscription [McCarthy, 1986, Lifschitz, 1985]. Moreover, like in these frameworks, the induced order on interpretations establishes a preference which favors models M which violate a smaller set $\Delta[M]$ of assumptions:

Lemma 4.2 *For two models M and M' of a theory T, if $\Delta[M] \subset \Delta[M']$, then M is preferred to M' ($M < M'$) in every prioritized preferential structure.*

In particular, then, preferred models violate a minimal set of assumptions:[3]

Lemma 4.3 *If M is a preferred model of a theory T in a given prioritized preferential structure, then M is* minimal *in $\Delta_{\mathcal{L}}$, i.e. there is no model M' of T such that $\Delta[M'] \subset \Delta[M]$.*

[3]Recall that a model of a default theory $T = \langle K, E \rangle$ with $K = \langle L, D \rangle$ is an interpretation that satisfies the sentences in both L and E.

The minimality of preferred models in prioritized preferential structures will be responsible for endowing the resulting form entailment, called *conditional entailment*, with features of non-monotonic logics absent from p-entailment and ϵ-entailment. For instance, given a theory $T = \langle K, E \rangle$, with a single default $p \rightarrow \delta$ and a body of evidence $E = \{p, q\}$, there will be models of T which violate no assumption at all. Thus, the lemma above states that the preferred models have to be among those models, and therefore, that the assumption δ will be true in all preferred models of T. As a result, δ will be conditionally entailed by T, though δ is not p-entailed or ϵ-entailed by T.

We will often refer to the sets of assumptions $\Delta[M]$ violated by an interpretation M, as the *gap* of the interpretation M. *Minimal models* are models with minimal gaps or, alternatively, models committed to a maximal set of assumptions.[4]

The order on interpretations induced from a priority relation on assumptions depends only on the interpretation gaps, and in particular, interpretations with identical gaps are not distinguished. In that regard, since models with an identical gap are models committed to the same set of assumptions, the induced ordering on interpretations can also be understood as an ordering on sets of assumptions. The notion of *classes* of models as collection of models committed to a common set of assumptions, will serve to make this view more explicit:

Definition 4.2 *A* class $\mathcal{C}$ *of a theory* T *with an associated gap* $\Delta[\mathcal{C}] = \Delta$, *represents the non-empty collection of models* M *of* T *such that* $\Delta[M] \subseteq \Delta$.

Thus a class $\mathcal{C}$ with gap $\Delta = \Delta[\mathcal{C}]$ includes all the models which validate all the assumptions not mentioned in Δ. In particular, we will say that the class $\mathcal{C}$ is *minimal*, iff for every model M in $\mathcal{C}$, $\Delta[M] = \Delta$. A minimal class is thus a collection of minimal models. We also say that a proposition p *holds in a class* $\mathcal{C}$ iff p holds in every model in $\mathcal{C}$. Proof-theoretically, this is equivalent to the existence of a classical derivation of p from the sentences in T and assumptions not in the gap of $\mathcal{C}$. Since preferred models are guaranteed to be minimal (lemma 4.3) and the induced preference relations do not distinguish between models with identical gaps, such preference relations can be usefully regarded as selecting, among the minimal classes of a theory T, the preferred ones.

[4]This notion of minimality is not the conventional one. Usually minimal models are defined in terms of the *extension* of some predicates (e.g. [McCarthy, 1980]), rather than in terms of the truths of some literals. We will discuss these issues in more detail in section 4.4.

In the rest of this section we will focus on three related issues: the well-foundedness and admissibility conditions of prioritized preferential structures, the constraints on admissible assumption priorities, and the form of entailment defined by admissible prioritized preferential structures.

The condition of well-foundness was introduced in section 3.2 as a condition that guarantees consistency and compliance with rules 1–5 of **P**. We said that a *default theory* T is *well-founded* in a preferential model structure $\langle \mathcal{I}, < \rangle$ when for every non-preferred model M of T in $\mathcal{I}$, there is a preferred model M' in $\mathcal{I}$ such that $M' < M$. Likewise, we said that a *preferential model structure* π is *well-founded* relative to a background context K when every theory $T = \langle K, E \rangle$ is well founded in π.

These notions could be easily generalized to prioritized preferential structures by stipulating that a structure $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$ is well-founded relative to a background context K when the embedded preferential model structure $\langle \mathcal{I}, < \rangle$ is. Such an approach, however, would be unnecessarily restrictive. Indeed, if the underlying language $\mathcal{L}$ is first order, for most prioritized preferential structures $\pi = \langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$ and background contexts K of interest, it will be possible to construct some theory $T = \langle K, E \rangle$, which is not well-founded in π, rendering every such π structure not well-founded.[5]

For that reason we will abandon here the notion of well-founded *prioritized preferential structures*, and restrict ourselves to the weaker notion of well-founded *theories*. A theory T is *well-founded in a structure* $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$, when T is well-founded in the embedded p-structure $\langle \mathcal{I}, < \rangle$, and is simply *well-founded* when it is well-founded in *in every prioritized structure.*

Definition 4.3 *A theory T is* well-founded *when it is well-founded in every prioritized preferential structure; namely, for every structure $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$ and every model M of T, there exists a preferred model M' of T, such that $M' < M$ or $M' = M$.*

Actually, even this notion is more restrictive that it needs to be. Still, as we will show below, it comprises a large class of theories which includes all the theories we will consider.

[5]Indeed, since preferred models in any prioritized preferential structure must be minimal, it is sufficient for T to be a theory which lacks minimal models, in the sense defined above (see, for instance, [Etherington, 1988, pp. 117]).

In order to guarantee that theories under consideration are well-founded in this sense, it will be convenient to show that they belong to a class of theories which we call *bound*. Bound theories are defined in terms of the *conflict sets* they give rise to and are always well-founded. If Δ stands for a set of assumptions and $\neg\Delta$ stands for the negation of the conjoin of Δ, conflict sets are defined as follows [Reiter, 1987b]:

Definition 4.4 *A set of assumptions* Δ *constitute a* conflict set *in a context* $T = \langle K, E \rangle$ *iff* $E \models_K \neg\Delta$.

Conflict sets are thus sets of assumptions which cannot all hold in a given context. The *minimal* conflict sets in a given context T contain assumptions which 'compete' with each other. On the other hand, assumptions which do not belong to any minimal conflict set are not 'questioned', and therefore, are guaranteed to hold in every minimal model of T. We call these assumptions *free* in T, and distinguish them from *bound* assumptions as follows:[6]

Definition 4.5 *An assumption which belongs to a minimal conflict set in* T *is* bound *in* T*; otherwise, it is* free *in* T.

For instance, the assumption that "Tim is a bird, then Tim flies" is free in a context in which all the evidence is that Tim is a bird, and it is bound in a context in which Tim is known to be a penguin. In most theories of interest most assumptions are free, with the exception of a small number of *bound* assumptions which point to possible 'abnormalities.' We call these theories, *bound* theories:

Definition 4.6 *A theory* T *is* bound *if it gives rise to a finite number of bound assumptions.*

Interestingly, this property is *sufficient* to make a theory *well-founded:*

Lemma 4.4 *Bound default theories are well-founded.*

[6]The complement of the assumptions that we call *free* are essentially what Gelfond *et al.* [1986] call *free for negation.*

This result thus guarantees that as long as we are dealing with theories which give rise to a finite number of possible 'abnormalities,' we are in a safe terrain: every model will be or will be outranked by a preferred model.

Having defined bound and well-founded theories, we are now in a position to specify the space of prioritized preferential structures $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$ which are *admissible* with a given background context. These structures will correspond to those which are induced from priority orderings '$\prec$' which obey with certain conditions sketched in section 2.5. We said then that a set of assumption Δ is in conflict with a default $p \rightarrow \delta$ in K, if Δ constitutes an argument against δ in the context $\langle K, \{p\} \rangle$, i.e. if $p, \Delta \vdash_K \neg\delta$ and $p \not\vdash_K \neg\Delta$. If so, we said that δ *directly dominates (d-dominates)* the set Δ and any superset of it. The intuition was that by writing a default $p \rightarrow \delta$ the user intends δ to be true in the preferred models of p, and thus, s/he implicitly regards the violation of one of the assumptions in Δ as less important than the violation of δ. *Admissible priority orderings* capture such intuition as follows:

Definition 4.7 *A priority order '$\prec$' over $\Delta_\mathcal{L}$ is* admissible *with a background context K iff every set Δ of assumptions directly dominated (d-dominated) by an assumption δ contains an assumption δ' such that $\delta' \prec \delta$.*

A similar intuition underlies the definition of admissible *default rankings* in section 3.5. Indeed, if K is a *pure* assumption based default theory,[7] the priority ordering '$\prec$' defined as $\delta_1 \prec \delta_2$ iff $\sigma(p_1 \rightarrow \delta_1) < \sigma(p_2 \rightarrow \delta_2)$ for defaults $p_1 \rightarrow \delta_2$ and $p_2 \rightarrow \delta_2$ in K and an admissible *default ranking* σ, will be admissible with K. Admissible priority orderings, however, are not admissible default rankings in general because priority ordering, unlike default rankings, are *partial orders*.[8]

Example 4.1 Admissible priority orderings, like default rankings, normally encode priorities which reflect the 'specificity' of defaults. Consider, for instance, a background context K encoding two defaults $\mathtt{p} \rightarrow_1 \mathtt{q}$ and $\mathtt{p} \wedge \mathtt{r} \rightarrow_2 \neg\mathtt{q}$. Namely, K contains the sentences $\mathtt{p} \wedge \delta_1 \Rightarrow \mathtt{q}$ and $\mathtt{p} \wedge \mathtt{r} \wedge \delta_2 \Rightarrow \neg\mathtt{q}$, and the defaults $\mathtt{p} \rightarrow \delta_1$

[7] $K = \langle L, D \rangle$ is *pure* if it is the translation of a background K' which does not contain assumption predicates (see section 2.1). Namely, assumptions predicates δ_i only occur in L in sentences of the form $p(x) \wedge \delta_i(x) \Rightarrow q(x)$ for default schemas $p(x) \rightarrow q(x)$ in K'.

[8] The advantage of focusing on partial as opposed to total orders is that the former provide a more concise representation of partial information. Often we will find background contexts which can be characterized in terms of a *single* admissible priority orderings, but which would otherwise require *multiple* admissible rankings.

and $\mathtt{p} \wedge \mathtt{r} \rightarrow \delta_2$. No set of assumptions is in conflict with the default $\mathtt{p} \rightarrow \delta_1$, and thus, $\mathtt{p} \rightarrow \delta_1$ does not impose any constraint on admissible priority orderings. On the other hand, the set $\{\delta_1\}$ is in conflict with the default $\mathtt{p} \wedge \mathtt{r} \rightarrow \delta_2$, as the relation $\mathtt{p} \wedge \mathtt{r}, \delta_1 \mid\!\!\sim_K \neg\delta_2$ holds, but $\mathtt{p} \wedge \mathtt{r} \mid\!\!\sim_K \neg\delta_1$ does not. As a result, a priority ordering '$\prec$' will be a admissible with K iff the relation $\delta_1 \prec \delta_2$ holds. In other words, in any admissible priority ordering the assumption δ_2 corresponding to the 'more specific' default $\mathtt{p} \wedge \mathtt{r} \rightarrow \delta_2$, will have a higher priority than the assumption δ_1 corresponding to the 'less specific' default $\mathtt{p} \rightarrow \delta_1$.

We will also refer to admissible priority orderings as *conditional* orderings, as the constraints they obey are a result of interpreting defaults of the form $p \rightarrow \delta$, conditionally; namely, assertions δ bound to a particular context p. The *admissible* prioritized preferential structures are the structures induced from admissible priority orderings:

Definition 4.8 *A prioritized preferential structure $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$ is* admissible *with a background $K = \langle L, D \rangle$, iff the priority ordering '$\prec$' is admissible with K, and every theory $T = \langle K, \{p\} \rangle$ for a default $p \rightarrow \delta$ in K is bound.*

Conditional entailment is defined in terms of the preferred models of the *admissible* prioritized structures:

Definition 4.9 *A proposition q is* conditionally entailed (cd-entailed) *by a default theory $T = \langle K, E \rangle$, when q holds in all the preferred models of T of every prioritized preferential structure admissible with K.*

We also refer to the preferred models of $T = \langle K, E \rangle$ in some admissible prioritized structure admissible as the *conditional models* of T, and similarly, to the preferred classes of T, as the *conditional classes* of T. This terminology will be particularly useful in chapter 5, when we study preference relations based on *causal* rather than *conditional* considerations.

There are some important differences between the definition of conditional entailment and the definition of preferential entailment in chapter 3. Even though every prioritized preferential structure $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$ embeds a preferential model structure $\langle \mathcal{I}_\mathcal{L}, < \rangle$, the admissibility conditions on both structures are very different. The admissibility of p-structures amounts, essentially, to a restriction on the

preferred models of the theories $T = \langle K, \{p\}\rangle$ for defaults $p \rightarrow \delta$ in K. Prioritized preferential structures, on the other hand, capture these constraints by means of the restriction on the priority orderings on assumptions (see below), but go beyond that by rewarding models which violate smaller sets of assumptions. In that way conditional entailment is able to combine the features of conditional interpretation of defaults with the features common to traditional non-monotonic logics.

There are two other differences worth pointing out. Admissible prioritized structures $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec\rangle$, unlike admissible preferential structures $\langle \mathcal{I}, <\rangle$, comprise the whole set of interpretations over the underlying language $\mathcal{L}$. This is needed to guarantee the minimality of preferred models; a notion which plays no role in the context of p-entailment. On the other hand, we do not require admissible prioritized structures to be well-founded as we demanded from admissible preferential structures. We have replaced this requirement by the more reasonable one of requiring the *theories* $T = \langle K, \{p\}\rangle$ for defaults $p \rightarrow \delta$ in K to be well-founded, and in particular, bound.[9]

The latter technical differences between admissible prioritized structures and admissible preferential structures makes a general comparison between preferential entailment and conditional entailment difficult. However, if we restrict ourselves to *finite propositional languages*, the following result is obtained:

Theorem 4.1 *A theory T preferentially entails a proposition p only if T conditionally entails p.*

Furthermore, since preferential entailment and ϵ-entailment coincide for finite propositional languages, the same subsumption relation applies to ϵ-entailment.

For the subsumption relations to be meaningful, though, we also need to guarantee that while T conditionally entails p, it does not entail $\neg p$. Namely, we need to show that a p-consistent background context, is also cd-consistent, where a background K is *cd-consistent* when it admits an admissible prioritized preferential structure. Indeed, when the background K is *pure* (see page 70), the following result holds for *finite propositional languages:*

Theorem 4.2 *A background context K is p-consistent only if K is cd-consistent.*

[9]This condition could be also relaxed in principle; though it is sufficiently ample for our purposes.

Theorems 4.1 and 4.2 thus show that conditional entailment captures the patterns of inference sanctioned by preferential entailment and ϵ-entailment. A simple example further reveals that conditional entailment goes well beyond them. Indeed, while preferential entailment and ϵ-entailment are *semi-monotonic* (non-monotonic in the evidence set but *monotonic* in the background context; theorem 2.2) conditionally entailment is fully *non-monotonic*.

Example 4.2 Let us consider a background K with a sentence $\mathtt{p} \wedge \delta_1 \Rightarrow \mathtt{q}$ and a default $\mathtt{p} \rightarrow \delta_1$, and a default theory $T = \langle K, \{\mathtt{p}\}\rangle$. First of all, note that T is a bound theory; as a matter of fact, T does not give rise to any conflict set, and thus all assumptions are free in T. The minimal models of T thus falsify no assumption and, therefore, all belong to a unique minimal class $\mathcal{C}$ of models with an empty gap. Lemma 4.3 guarantees then that $\mathcal{C}$ is the only preferred class of T and, hence that the assumption δ_1 holds in every preferred model of p in K. Moreover, since every model that satisfies p, δ_1, and the sentences in K must also satisfy q, then q is conditionally entailed by p as well.

The same reasoning applies to a context in which $\neg$q is observed instead of p. That is, the preferred models of the theory $T' = \langle K, \{\neg\mathtt{q}\}\rangle$ will also correspond to the empty gap interpretations that satisfy T', and thus, both sentences δ_1 and $\neg$p will be conditionally entailed by T'. Moreover, by similar arguments, both conclusions remain in the presence of an additional piece of evidence e, though they would have to be retracted if K is is augmented by a sentence $\mathtt{e} \Rightarrow \delta_1$.

The example shows that conditional entailment is able to capture reasoning patterns that involve 'default modus tollens' and conditional independence assumptions which escape ϵ-entailment and p-entailment. The key feature responsible for the additional power of conditional entailment is the minimality of preferred models. In this way, conditional entailment not only captures the *context sensitivity* of defaults in a way analogous to ϵ-entailment and p-entailment, but also the more conventional property by which the connection between default antecedent and consequent is preserved in the absence of conflicting evidence.

The examples below illustrate other features of conditional entailment. The reader familiar with prioritized circumscription [McCarthy, 1986] will notice the similarities. It should be kept in mind, however, that while prioritized circumscription accepts priorities from the user, conditional entailment *extracts* the priorities

automatically from the defaults present in the background context.[10]

We will find useful to write $\Delta \prec \delta$ as an abbreviation of "there exists a δ' in Δ such that $\delta' \prec \delta$." The admissibility of the priority order '$\prec$' with respect to K, can thus be expressed as the condition that if δ d-dominates Δ then the relation $\Delta \prec \delta$ must hold.

Example 4.3 (Strict Specificity) Let us consider a background context $K = \langle L, D \rangle$ with sentences:

$$\mathtt{b}(x) \wedge \delta_1(x) \Rightarrow \mathtt{f}(x)$$
$$\mathtt{p}(x) \wedge \delta_2(x) \Rightarrow \neg\mathtt{f}(x)$$
$$\mathtt{p}(x) \Rightarrow \mathtt{b}(x)$$
$$\mathtt{r}(x) \Rightarrow \mathtt{b}(x)$$

and a pair of defaults $\mathtt{b}(x) \rightarrow \delta_1(x)$ and $\mathtt{p}(x) \rightarrow \delta_2(x)$ (fig. 4.2).

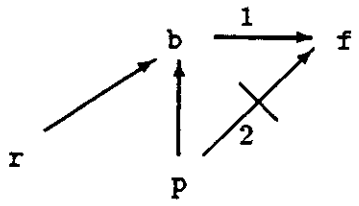


Figure 4.2: Strict specificity

A priority ordering '$\prec$' is admissible with K if the relation $\Delta \prec \delta_i(a)$ holds for any minimal assumption set Δ d-dominated by $\delta_i(a)$, for $i = 1, 2$ and any term a in the language. The assumption $\delta_1(a)$ does not d-dominate any set Δ so there are no admissibility constraints originated from the default schema $\mathtt{b}(x) \rightarrow \delta_1(x)$. On the other hand, the assumption $\delta_2(a)$ d-dominates a single minimal set of assumptions $\Delta = \{\delta_1(a)\}$, as for any ground term a, we have $\mathtt{p}(a), \Delta \mid\!\!\sim_K \neg\delta_2(a)$, while $\mathtt{p}(a) \not\mid\!\!\sim_K \neg\Delta$. Thus, a priority order '$\prec$' will be admissible with K iff the relation $\delta_1(a) \prec \delta_2(a)$ holds for every ground term a. We also write in these cases simply $\delta_1(x) \prec \delta_2(x)$.

Provided with this implicit characterization of the prioritized preferential structures admissible with K, we can now turn to analyze the propositions conditionally

[10]The relation between conditional entailment and prioritized circumscription will be discussed in some detail in section 4.4.

entailed in the different contexts of interest. For an individual Tim (t), the preferred models of b(t) in K are once again, the models which violate no assumption. As a result, both the assumptions $\delta_1(\mathtt{t})$ and $\delta_2(\mathtt{t})$ are conditionally entailed by T, as are the propositions f(t) and $\neg$p(t). A different scenario arises, however, if we consider the evidence p(t) instead of b(t). In that case, every interpretation satisfying the evidence and the background context is forced to render one of assumptions $\delta_1(\mathtt{t})$ or $\delta_2(\mathtt{t})$ false. Thus, two classes of minimal models arise: a class $\mathcal{C}_{\{1\}}$ whose models M_1 have a gap $\Delta[M_1] = \{\delta_1\}$, and a class $\mathcal{C}_{\{2\}}$ whose models M_2 have a gap $\Delta[M_2] = \{\delta_2\}$. However, the former class is preferred to the latter as for any such models M_1 and M_2, the relation $M_1 < M_2$ in every admissible prioritized structure. Indeed, $\Delta[M_2] - \Delta[M_1] = \{\delta_2(\mathtt{t})\}$, $\Delta[M_1] - \Delta[M_2] = \{\delta_1(\mathtt{t})\}$, and the relation $\delta_1(\mathtt{t}) \prec \delta_2(\mathtt{t})$ holds for every admissible priority ordering. It follows then, that $\mathcal{C}_{\{1\}}$ represents the class of preferred models of p(t) in K, and therefore, that the propositions $\delta_2(\mathtt{t})$ and $\neg$f(t), as expected, are conditionally entailed by p(t) in K. Similar conclusions, indeed, are legitimized by p-entailment and ϵ-entailment.

Consider now the scenario in which the target context is enhanced with the information that Tim is also a red bird, i.e. $T' = \langle K, E' \rangle$, where $E' = \{\mathtt{p}(\mathtt{t}), \mathtt{r}(\mathtt{t})\}$. In this case, neither ϵ-entailment nor p-entailment constrain the preferred models of T'. In cd-entailment, on the other hand, we are guaranteed that the preferred models of T' are minimal, and therefore, that they belong to one of the minimal classes $\mathcal{C}_{\{1\}}$ and $\mathcal{C}_{\{2\}}$, where $\mathcal{C}_I$ stands for the class of models M of T' with a gap $\Delta[M] = \{\delta_i(\mathtt{t}) \mid i \in I\}$. However, as we showed above, models in $\mathcal{C}_{\{1\}}$ are preferred to models in $\mathcal{C}_{\{2\}}$.[11] As a result, in agreement with the irrelevance account given in chapter 2, the assumption $\delta_2(\mathtt{t})$ and the proposition $\neg$f(t) are conditionally entailed by T'.

The example above illustrates a background context where every admissible priority ordering '$\prec$' must include all tuples of the form $\langle \delta_1(a), \delta_2(a) \rangle$, for ground terms a in the language. Priority orderings may also include additional tuples, e.g. $\langle \delta_1(\mathtt{a}), \delta_2(\mathtt{b}) \rangle$, but those tuples are not necessary for the ordering to be admissible. We will say that an admissible priority ordering '$\prec$' is *minimal* when there in no admissible priority ordering '$\prec'$' which includes a proper subset of the tuples in '$\prec$.' In other words, a minimal admissible priority ordering is an ordering from which no set of tuples can be removed without violating the admissibility conditions. For instance, in the example above, there is a *single* minimal admissible

[11] Note that these classes do not contain the same interpretations as in the context T above. Still, they possess the same gaps, and the preference relation on classes exclusively depends on such gaps.

ordering which includes all the tuples of the form $\langle \delta_1(x), \delta_2(x) \rangle$ and nothing else. Later we will see background contexts which give rise to *multiple* minimal admissible priority orderings. In any case, due to the fact that minimal priority orderings contain only the strictly necessary relationships, it is natural to ask whether conditional entailment can be computed by restricting attention to *minimal* admissible priority orderings as opposed to *all* admissible priority orderings. The answer is yes. Indeed, if we can obtain the admissible priority ordering '$\prec$' by deleting certain tuples from the admissible priority ordering '$\prec'$,' the preferred models in the structure $\langle \mathcal{I}_\mathcal{L}, <', \cdot_\mathcal{L}, \prec' \rangle$ will be a subset of the preferred models of the structure $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$. Thus, if we say that an admissible prioritized preferential structure $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$ is *minimal* if the relation '$\prec$' is a minimal admissible priority ordering, the following result holds:

Lemma 4.5 *A proposition q is* conditionally entailed (cd-entailed) *by a default theory $T = \langle K, E \rangle$, when q holds in all preferred models of T of every* minimal *prioritized preferential structure admissible with K.*

In the example above, we can thus compute conditional entailment by considering a *single* structure $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$, where the priority ordering is such that $\delta \prec \delta'$ holds iff $\delta = \delta_1(a)$ and $\delta' = \delta_2(a)$ for some ground term a in the language. Often, however, multiple structures will need to be considered (see example 4.6 below).

Example 4.4 (Cycles) In this example we illustrate the prioritization associated with a cyclic default inheritance network. The background context $K = \langle L, D \rangle$ encodes the following default schemas (fig. 4.3):

$$\mathtt{c}(x) \rightarrow_1 \mathtt{u}(x)$$
$$\mathtt{u}(x) \rightarrow_2 \mathtt{a}(x)$$
$$\mathtt{a}(x) \rightarrow_3 \neg\mathtt{u}(x)$$

Namely, for each such expression $p(x) \rightarrow_i q(x)$, K contains a default schema $p(x) \rightarrow \delta_i(x)$ and a sentence $p(x) \wedge \delta_i(x) \Rightarrow q(x)$. The defaults above can be read as stating that "most people sitting in the class are university students," "most university students are adults," and "most adults are not university students."

In order to determine the space of prioritized preferential structures admissible with K, we need to look at the pattern of dominances it gives rise to. In this

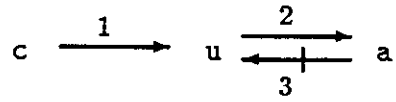


Figure 4.3: A cyclic inheritance hierarchy

background context, there are only two such patterns. First, assumptions of the form $\delta_2(a)$ d-dominate the set $\{\delta_3(a)\}$. This can be inferred from the presence of the path $\mathtt{u} \rightarrow \mathtt{a} \not\rightarrow \mathtt{u}$ in the figure, and more precisely, from the expressions $\mathtt{u}(a), \delta_3(a) \mid\!\sim_K \neg\delta_2(a)$ and $\mathtt{u}(a) \not\mid\!\sim_K \neg\delta_3(a)$. As a result, any admissible priority ordering '$\prec$' must be such that $\delta_3(x) \prec \delta_2(x)$ holds. Secondly, the assumption $\delta_1(a)$ d-dominates the set $\{\delta_2(a), \delta_3(a)\}$, and thus every admissible priority ordering must also satisfy $\delta_2(a) \prec \delta_1(a)$ or $\delta_3(a) \prec \delta_1(a)$, a disjunction which we abbreviate as $\{\delta_2(a), \delta_3(a)\} \prec \delta_1(a)$.

Note that if the priority order is such that $\delta_2(a) \prec \delta_1(a)$ holds, by the transitivity of '$\prec$', it must also be the case that the relation $\delta_3(a) \prec \delta_1(a)$ holds. Moreover, since either $\delta_2(a)$ or $\delta_3(a)$ *must* have a lower priority than $\delta_1(a)$, the relation $\delta_3(a) \prec \delta_1(a)$ must hold even if $\delta_2(a) \not\prec \delta_1(a)$. As a result, there is a single *minimal* admissible priority ordering '$\prec$' which only satisfy $\delta_3(a) \prec \delta_1(a)$ and $\delta_3(a) \prec \delta_2(a)$, for ground terms a in the language.

We can illustrate the behavior of conditional entailment under such a prioritization by considering a context $T = \langle K, E\rangle$, where $E = \{\mathtt{c}(\mathtt{k}), \mathtt{a}(\mathtt{k})\}$. Namely, we know that an adult, say Ken, is sitting in the class; we want to know whether he is likely to be a university student, $\mathtt{u}(\mathtt{k})$. First, the context T gives rise to two minimal classes of models: a class $\mathcal{C}_{\{1\}}$ of models which only violate the assumption $\delta_1(\mathtt{k})$ ('Ken is in the class, then he is a university student'), and a class $\mathcal{C}_{\{3\}}$ of models which only violate the assumption $\delta_3(\mathtt{k})$ ('Ken is an adult, then he is not a university student'). However, the latter class is preferred to the former one, as for any models $M_1 \in \mathcal{C}_{\{1\}}$ and $M_3 \in \mathcal{C}_{\{3\}}$, $\Delta[M_1] - \Delta[M_3] = \{\delta_1(\mathtt{k})\}$, $\Delta[M_3] - \Delta[M_1] = \{\delta_3(\mathtt{k})\}$, and $\delta_3(\mathtt{k}) \prec \delta_1(\mathtt{k})$. As a result, and since $\delta_3(\mathtt{k})$ is the only assumption violated in the preferred class $\mathcal{C}_{\{3\}}$, the assumption $\delta_1(\mathtt{k})$ is conditionally entailed and so is the target proposition $\mathtt{u}(\mathtt{k})$. The same result follows indeed from p-entailment and ϵ-entailment. However, while the same derivation remains in conditional entailment in the presence of irrelevant information, e.g. Ken is blond, it is no longer legitimate in either p-entailment or ϵ-entailment.

Example 4.5 (Default Specificity) We consider now a slightly different background context $K = (L, D)$, representing the hierarchy depicted in figure 4.4. For simplicity, we deal the propositional defaults:

$$\mathtt{a} \rightarrow_1 \mathtt{w}$$
$$\mathtt{u} \rightarrow_2 \neg\mathtt{w}$$
$$\mathtt{u} \rightarrow_3 \mathtt{a}$$
$$\mathtt{f} \rightarrow_4 \mathtt{a}$$

As usual we assume that each such expression $p \rightarrow_i q$ abbreviates the pair comprised by the sentence $p \wedge \delta_i \Rightarrow q$ and the default $p \rightarrow \delta_i$. The defaults above can be understood as expressing "most adults work," "most university students do not work," "most university students are adults," and "most fans of Frank Sinatra are adults," respectively.

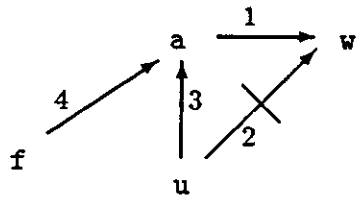


Figure 4.4: Default Specificity

There are two relevant dominance relations in this background context. First, the assumption δ_2 d-dominates the set $\Delta = \{\delta_1, \delta_3\}$ as Δ constitutes an argument against the default $\mathtt{u} \rightarrow \delta_2$, i.e. $\mathtt{u}, \Delta \mid\!\sim_K \neg\delta_2$ holds, while $\mathtt{u} \mid\!\sim_K \neg\Delta$ does not. Likewise, the assumption δ_3 d-dominates the set $\{\delta_1, \delta_2\}$. Thus, any priority ordering '$\prec$' admissible with K must be such that both relations $\{\delta_1, \delta_3\} \prec \delta_2$ and $\{\delta_1, \delta_2\} \prec \delta_3$ are satisfied. Moreover, due to the asymmetric and transitive character of priority orderings, such constraints can be further simplified to $\delta_1 \prec \delta_2$ and $\delta_1 \prec \delta_3$. To show this is the case, let us assume $\delta_2 \prec \delta_3$. Then, by the asymmetry of the priority order we must have $\delta_3 \not\prec \delta_2$, and therefore, from the constraints above, $\delta_1 \prec \delta_2$. On the other hand, if $\delta_2 \not\prec \delta_3$ and $\delta_1 \not\prec \delta_2$, the constraints above imply $\delta_1 \prec \delta_3$ and $\delta_3 \prec \delta_2$ in contradiction with the transitivity of '$\prec$'. Thus, either if $\delta_2 \prec \delta_3$ holds or not, the relation $\delta_1 \prec \delta_2$ must hold. On similar arguments, we can infer that the relation $\delta_1 \prec \delta_3$ must hold as well.

With these space of admissible priority orderings, let us consider first a context $T = (K, E)$, with $E = \{\mathtt{f}\}$. Since there is an interpretation that satisfies T and

every assumption in the language, the single preferred class in every prioritized structure admissible with K turns out to be the class of models of T which violate no assumption. In particular, the assumptions δ_1 and δ_4 are conditionally entailed by T and so the propositions $\mathtt{a}$ and $\mathtt{w}$. Note that these inferences involve default chaining, a pattern not sanctioned by ϵ-entailment or by p-entailment.

A different situation arises, however, if the proposition $\mathtt{u}$ is also observed. The context $T' = \langle K, E' \rangle$, with $E' = \{\mathtt{f}, \mathtt{u}\}$, gives rise to three minimal classes: $\mathcal{C}_{\{1\}}$, $\mathcal{C}_{\{2\}}$ and $\mathcal{C}_{\{3,4\}}$, where $\mathcal{C}_I$, as usual, stands for the class of models M of T' such that $\Delta[M] = \{\delta_i : i \in I\}$. However, any model M in $\mathcal{C}_{\{1\}}$ is preferred to any model M' in $\mathcal{C}_{\{2\}}$ and any model M'' in $\mathcal{C}_{\{3,4\}}$. Indeed, $\Delta[M] - \Delta[M'] = \Delta[M] - \Delta[M''] = \{\delta_1\}$, $\Delta[M'] - \Delta[M] = \{\delta_2\}$, and $\Delta[M''] - \Delta[M] = \{\delta_3, \delta_4\}$, while the relations $\delta_1 \prec \delta_2$ and $\delta_1 \prec \delta_3$ hold in every admissible priority order. Hence $\mathcal{C}_{\{1\}}$ is the preferred class of T', and therefore, all assumptions with the exception of δ_1 are conditionally entailed by T', and are the propositions $\mathtt{a}$ and $\neg\mathtt{w}$.

The previous examples illustrate background contexts which can be characterized in terms of a single minimal admissible priority ordering. The next example illustrates a background context K which results in multiple ones. Namely, K dictates *disjunctive* constraints of the form $\Delta \prec \delta$, where Δ is a non-singleton assumption set which cannot be further simplified.

Example 4.6 Let K represent the hierarchy depicted in fig. 4.5 in propositional form:

$$\mathtt{a} \rightarrow_1 \mathtt{b}$$
$$\mathtt{a} \rightarrow_2 \mathtt{d}$$
$$\mathtt{b} \rightarrow_3 \mathtt{c}$$
$$\mathtt{c} \rightarrow_4 \neg\mathtt{d}$$

In order to determine the constraints admissible priority orderings must obey, we have to identify first the relevant dominance patterns. There are two such patterns: the assumption δ_1 d-dominates the assumption set $\{\delta_2, \delta_3, \delta_4\}$, and the assumption δ_2 d-dominates the assumption set $\{\delta_1, \delta_3, \delta_4\}$. Such relations are a result of the fact that $\mathtt{a} \mid\!\!\sim_K \neg(\delta_1 \wedge \delta_2 \wedge \delta_3 \wedge \delta_4)$ holds, while neither $\mathtt{a} \mid\!\!\sim_K \neg(\delta_2 \wedge \delta_3 \wedge \delta_4)$, nor $\mathtt{a} \mid\!\!\sim_K \neg(\delta_1 \wedge \delta_3 \wedge \delta_4)$ does. Thus, every priority ordering admissible with K must be such that both relations $\{\delta_2, \delta_3, \delta_4\} \prec \delta_1$ and $\{\delta_1, \delta_3, \delta_4\} \prec \delta_2$ are satisfied. Moreover, from these two constraints and the fact that priority orderings are asymmetric

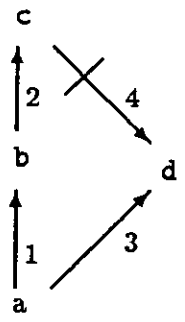


Figure 4.5: Disjunctive constraints

and transitive, it is possible to show as above that every admissible ordering must also comply with the simplified constraints $\{\delta_3, \delta_4\} \prec \delta_1$ and $\{\delta_3, \delta_4\} \prec \delta_2$. These constraints, however, cannot be simplified further.

Let us consider now a body of evidence $E = \{\mathtt{a}\}$. The theory $T = \langle K, E \rangle$ gives rise to four classes $\mathcal{C}_{\{i\}}$ of minimal models, each one with an associated gap $\{\delta_i\}$, $i = 1, \ldots, 4$. We show first that the preferred models of T in any prioritized preferential structure $\tau = \langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$ admissible with K are contained among those of $\mathcal{C}_{\{3\}}$ and $\mathcal{C}_{\{4\}}$. Let M_i be a model in $\mathcal{C}_{\{i\}}$, for $i = 1, \ldots, 4$ and assume that the relation $\delta_3 \prec \delta_1$ holds. Then, since $\Delta[M_3] - \Delta[M_1] = \{\delta_3\}$ and $\Delta[M_1] - \Delta[M_3] = \{\delta_1\}$, M_3 must be preferred to M_1 in τ and, therefore, M_1 is not a preferred model of T in τ. Assume now otherwise, that $\delta_3 \not\prec \delta_1$. Then from the constraint $\{\delta_3, \delta_4\} \prec \delta_1$ above, the relation $\delta_4 \prec \delta_1$ must be true. By similar arguments, it follows then that M_4 is preferred to M_1 in τ and therefore, that M_1, again, is not a preferred model of T in τ. Replacing M_1 by M_2, we obtain similarly that M_2 is not a preferred model of T either. Furthermore, since neither δ_3 has priority higher than δ_4, nor vice versa, no class among $\mathcal{C}_{\{3\}}$ and $\mathcal{C}_{\{4\}}$ is preferred to the other, and thus, both turn out to be the preferred classes of T. As a result, the assumptions δ_1 and δ_2 are conditionally entailed by T, and so the propositions b and d.[12]

The examples above illustrate the power gained by focusing on the class of admissible prioritized preferential structures over the larger class of admissible preferential structures. Prioritized preferential structures combine the constraints

[12]Note, however, that against intuition, the proposition c is not conditionally entailed by a since c does not hold in $\mathcal{C}_{\{3\}}$. More about this problem in chapter 5.

inherent to admissible preferential structures with the minimality constraints typical of more traditional non-monotonic formalisms. However, as the examples also show, uncovering the propositions which are conditionally entailed by a given default theory is far from a trivial task. First we need to determine the conditions to be met by the family of admissible priority orders, then identify the minimal classes, and only then can we uncover the set of preferred models and the conclusions they legitimize. In the next section, we will develop a proof theory which will enable us to bypass this process. Such a proof theory will permit us uncovering the conditionally entailed propositions by purely syntactic means.

4.2.2 Proof Theory

In this section we will focus on the development of syntactic criteria to determine when a proposition is conditionally entailed (cd-entailed) by a given default theory. We start analyzing the assertability conditions of *assumptions*, i.e. sufficient conditions under which an assumption is guaranteed to hold in all preferred models of a particular theory. We continue refining theses syntactic conditions to arrive, at the end, to a *sound* and *complete* syntactic characterization of conditional entailment.

It will be useful to recall some terminology. A set Δ of assumptions constitutes an argument in a context $T = \langle K, E \rangle$ if Δ is logically consistent with T. Δ is an argument *for* p if $E, \Delta \models_K p$, and an argument *against* p if $E, \Delta \models_K \neg p$. If Δ is not logically consistent with T, then Δ is *conflict set* in T. *Two arguments are in conflict* when their union is a conflict set. Likewise, an assumption is *free* in T when it does not belong to any *minimal* conflict set in T, and is *bound* otherwise. Default theories T are *bound* when they give rise to a *finite* set of bound assumptions. All theories considered so far are bound and as the theories to which the syntactic account below applies.

The first condition for assertability is a simple consequence of the minimality of preferred models within the class of prioritized preferential structures.

Lemma 4.6 *If an assumption δ is free in T, then δ is conditionally entailed by T.*

A similar condition is both sound and complete for circumscriptive theories, provided that assumptions are identified with negative literals and that T includes the *unique names* and *domain closure* axioms [Gelfond and Przymusinska, 1986].

In the context of conditional entailment, however, such a condition is too weak. Often an assumption is bound in a context T and still is conditionally entailed. The "birds fly–penguins don't" (example 4.3) provides one such example. In the context in which Tim is known to be a penguin, the assumptions 'Tim flies, because it is bird' ($\delta_1(\mathtt{t})$) and 'Tim does not fly, because it is a penguin' ($\delta_2(\mathtt{t})$) are in conflict, and thus, both are bound. Still the latter assumption is conditionally entailed.

In order to capture these conclusions by syntactic means, we need to go beyond the minimality of preferred models to consider the constraints imposed by K on the admissible priority orderings. Indeed, the reason the assumption $\delta_2(\mathtt{t})$ holds in spite of being in conflict with $\delta_1(\mathtt{t})$ is because its priority is higher; that is, under every priority ordering admissible with K the relation $\delta_1(\mathtt{t}) \prec \delta_2(\mathtt{t})$ holds.

The assertability conditions below take these constraints into account. Recall that we write $\Delta \prec \delta$ as an abbreviation of the expression $\exists \delta' \in \Delta$ such that $\delta' \prec \delta$.

Lemma 4.7 *An assumption δ is conditionally entailed by a default theory $T = \langle K, E \rangle$ if for every argument Δ against δ in T and every priority ordering '$\prec$' admissible with K, the relation $\Delta \prec \delta$ holds.*

Note that it is sufficient to consider the *minimal* arguments Δ against δ; if the relation $\Delta \prec \delta$ holds, so will the relation $\Delta' \prec \delta$ for any superset of Δ.

This new condition provides a correct handling of the example above. In the context in which Tim is known to be a penguin, the set $\Delta = \{\delta_1(\mathtt{t})\}$ ("if Tim is a bird, then Tim flies") is the only (minimal) argument against $\delta_2(\mathtt{t})$ ("if Tim is a penguin, then Tim doesn't fly"), and since $\delta_2(\mathtt{tim})$ has a priority higher than $\delta_1(\mathtt{tim})$, lemma 4.7 permits us to derive $\delta_2(\mathtt{t})$, and so the proposition 'Tim does not fly.'

While lemma 4.7 refines lemma 4.6, it is not yet complete with respect to conditional entailment. This can be illustrated by converting the *strict* subsumption 'links' $\mathtt{p}(x) \Rightarrow \mathtt{b}(x)$ ('penguins are birds'), and $\mathtt{r}(x) \Rightarrow \mathtt{b}(x)$ ('red-birds are birds') by *default* subsumption 'links' $\mathtt{p}(x) \rightarrow_3 \mathtt{b}(x)$ and $\mathtt{r}(x) \rightarrow_4 \mathtt{b}(x)$, with associated assumption predicates δ_3 and δ_4 respectively (fig. 4.6).

Since the resulting structure is identical to that analyzed in example 4.5, by similar arguments it is possible to show that the assumption $\delta_4(\mathtt{t})$ is conditionally entailed by the context $T = \langle K, E \rangle$, with $E = \{\mathtt{p}(\mathtt{t}), \mathtt{r}(\mathtt{t})\}$. Yet, the conditions in lemma 4.7 do not authorize asserting $\delta_4(\mathtt{t})$ in T: $\Delta = \{\delta_1(\mathtt{t}), \delta_2(\mathtt{t})\}$ is an argument against $\delta_4(\mathtt{t})$ for which the relation $\Delta \prec \delta_4(\mathtt{t})$ does not hold.

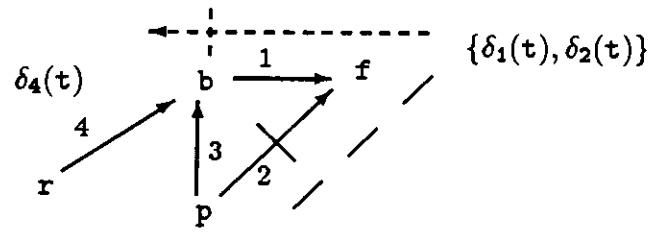


Figure 4.6: $\delta_4(t)$ is cd-entailed in spite of conflict with $\{\delta_1(t), \delta_2(t)\}$

Intuitively, the reason $\delta_4(t)$ is conditionally entailed by T in spite of the counterargument Δ, is that Δ contains an assumption $\delta_1(t)$ which is *defeated* in T. Namely, $\delta_1(t)$ which is in conflict with two 'better' assumptions $\delta_2(t)$ and $\delta_3(t)$. Thus, the latter two assumptions knock the argument Δ out, leaving the assumption $\delta_4(t)$ unchallenged.

In order to extend the assertability conditions of lemma 4.7 to handle such cases, we need to consider multiple conflicts at the same time. For that, some definitions will be handy.[13] We write below $\Delta' \prec \Delta$ as an abbreviation of the expression "for every δ in Δ, $\Delta' \prec \delta$" and assume a context $T = \langle K, E \rangle$.

Definition 4.10 *Given a priority ordering '$\prec$', an argument* Δ defeats *an argument* Δ' *if the two arguments are in conflict and the relation* $\Delta' \prec \Delta$ *holds. We say in that case that* Δ *is a* defeater *of* Δ'.

Definition 4.11 *An argument* Δ *is* protected *from a conflicting argument* Δ' *iff for every priority ordering admissible with* K, Δ *contains a defeater of* Δ'.

Intuitively, when an argument Δ is protected from a conflicting argument Δ' it means that Δ is a stronger argument than Δ', and that from the point of view of Δ, the conflicting argument Δ' can be ignored. When all such conflicting arguments can be so ignored, we say that Δ is a *stable:*

Definition 4.12 *An argument* Δ *is* stable *iff it is protected from every conflicting argument* Δ'.

[13]The vocabulary below is borrowed from argument-based approaches such as Loui's [1987a] and Pollock's [1987].

As suggested above, a stable argument is better than any of its competitors, and assumptions which belong to stable arguments are conditionally entailed:

Lemma 4.8 *If an assumption δ belongs to a stable argument Δ in T, then δ is conditionally entailed by T.*

The conditions in the previous lemma 4.7 can now be interpreted as requiring that the singleton argument $\{\delta\}$ be stable for the assumption δ to be assertable.

The notion of stable arguments is a very powerful one, sufficient to account for most inferences authorized by conditional entailment. In the example constructed above (fig. 4.6), for instance, $\Delta = \{\delta_2(\mathtt{t}), \delta_3(\mathtt{t}), \delta_4(\mathtt{t})\}$ is a stable argument in the context $T = \langle K, E\rangle$ with $E = \{\mathtt{p}(\mathtt{t}), \mathtt{r}(\mathtt{t})\}$, as the only minimal conflicting argument $\{\delta_1(\mathtt{t})\}$ is defeated by the subset $\Delta_s = \{\delta_2(\mathtt{t}), \delta_3(\mathtt{t})\}$ of Δ. As a result, lemma 4.8 authorizes us to conclude that any of the assumptions in Δ, and $\delta_4(\mathtt{t})$ in particular, are conditionally entailed by T.

The assertability conditions established by lemma 4.8 are powerful enough to account for all the examples we have considered so far. However, as the next example illustrates, they are yet not complete.

Example 4.7 Let us consider a background context K given by the sentences $p \wedge \delta_i \Rightarrow \neg\delta_3$, $i = 1, 2$, and the defaults $p \rightarrow \delta_i$, $i = 1, 2$. A priority ordering '$\prec$' will then be admissible with K iff $\delta_3 \prec \delta_1$ and $\delta_3 \prec \delta_2$. The context $T = \langle K, E\rangle$, with $E = \{p, \neg(\delta_1 \wedge \delta_2), \neg(\delta_3 \wedge \delta_4)\}$, gives rise to three minimal classes (fig. 4.7): $\mathcal{C}_{\{1,3\}}$, $\mathcal{C}_{\{2,3\}}$, and $\mathcal{C}_{\{1,2,4\}}$, where $\mathcal{C}_I$ stands for the set of models M of T such that $\Delta[M] = \{\delta_i : i \in I\}$. Furthermore, $\mathcal{C}_{\{1,3\}}$ and $\mathcal{C}_{\{2,3\}}$ are the preferred classes of T, and thus, the assumption δ_4 is conditionally entailed. However, δ_4 *does not* belong to any stable argument in T: neither argument among $\{\delta_4\}$, $\{\delta_1, \delta_4\}$ or $\{\delta_2, \delta_4\}$ is stable in T, and the assumption set $\{\delta_1, \delta_2, \delta_4\}$ is logically inconsistent with T. Thus, δ_4 is not derivable from the conditions in lemma 4.8.

The context T that corresponds to this example is depicted in fig. 4.7 (crossed links indicate incompatible pairs of assumptions). The problem is that while the assumption δ_4 cannot be protected from the conflicting assumption δ_3 by the 'better' assumptions δ_1 and δ_2 because $\Delta = \{\delta_1, \delta_2, \delta_4\}$ is inconsistent with T, the assumption δ_4 could be protected from δ_3 by the *disjunction* $\delta_1 \vee \delta_2$. That is, even though we cannot assert the sentence $\delta_1 \wedge \delta_2 \wedge \delta_4$, we would like to be able to test

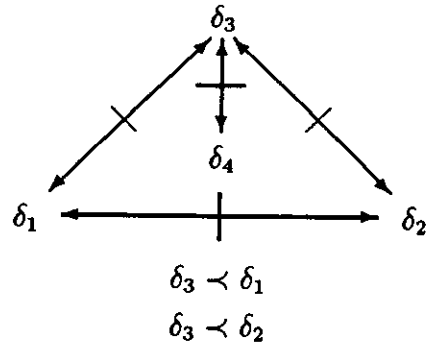


Figure 4.7: Beyond stable arguments

and assert the disjunctive sentence $(\delta_1 \wedge \delta_4) \vee (\delta_2 \wedge \delta_4)$. So, even when neither argument $\Delta_1 = \{\delta_1, \delta_4\}$ or $\Delta_2 = \{\delta_2, \delta_4\}$ is stable, because they conflict with the assumptions δ_2 and δ_1 respectively, their disjunction might. We will call such a collection of arguments $\{\Delta_1, \Delta_2\}$ which support the truth of δ_4 a *cover* for δ_4. Roughly, a *cover* C will be said to be *stable* when the *disjunction* of arguments in C — the sentence $(\delta_1 \wedge \delta_4) \vee (\delta_2 \wedge \delta_4)$ in this case — is assertable.

We make the notion of *stable covers* precise by refining first the conditions under which an argument is protected:

Definition 4.13 *An argument* Δ *is* strongly protected *from a conflicting argument* Δ' *if* Δ *is protected from* every conflicting argument Δ'' in Δ'.

For instance, if an argument Δ is protected from a conflicting argument Δ'_1 but is not protected from a conflicting argument Δ'_2, Δ will *not* be strongly protected from their union $\Delta'_1 + \Delta'_2$ even though Δ will be protected it. The distinction between the two notions is irrelevant for stable arguments which are *both* protected and strongly protected from every conflicting argument. As we will see, it is nonetheless needed for dealing with *disjunctive* arguments.

Let us refer to a collection of arguments as a *cover* — where a cover is to be understood as the disjunction of the arguments it contains — and let us generalize the notions of conflicts and protection to covers as follows:

Definition 4.14 *An argument* Δ *is* in conflict with a cover *if* Δ *is in conflict with*

every argument in the cover.

Definition 4.15 A cover is protected *from a conflicting argument* Δ *if the cover contains an argument* Δ' *which is* strongly protected *from* Δ.

The conditions under which a *cover* is stable then, are a straightforward generalization of the conditions under which an *argument* is stable:

Definition 4.16 *A* cover C *is* stable *iff it is protected from every conflicting argument* Δ.

Stable covers are thus a generalization of stable arguments, and if an argument Δ is stable so will be the singleton cover $\{\Delta\}$. So if we say that an *assumption belongs to a cover* when the assumption belongs to every argument in the cover, we will thus have assumptions which belong to stable covers which do not belong to stable arguments. The assumption δ_4 in the example 4.7 above, is one such case: δ_4 does not belong to any stable argument in the context T, and yet δ_4 belongs to the *stable cover* $\{\Delta_1, \Delta_2\}$. To prove this, let us first identify the arguments Δ' in conflict with *both* Δ_1 and Δ_2 in T. For Δ' to be in conflict with Δ_1, Δ' must contain one of the assumptions δ_2 or δ_3. Similarly, for Δ' to be in conflict with Δ_2, Δ' must contain one of the assumptions δ_1 or δ_3. Thus, if Δ' *does not* contain the assumption δ_3 it must contain both δ_1 and δ_2. However, this is not possible: such an assumption set is inconsistent with T and, therefore, it cannot be an argument. Hence, every argument Δ' in conflict with the cover $\{\Delta_1, \Delta_2\}$ must contain the assumption δ_3. Furthermore, only subsets of Δ' containing the assumption δ_3 can be in conflict with *either* Δ_1 or Δ_2. However, since both assumptions δ_1 and δ_2 are in conflict with δ_3 and have a higher priority than δ_3, it follows that *both* Δ_1 and Δ_2 are strongly protected from Δ', and therefore, that together, they constitute a stable cover.

As expected, the conditions of lemma 4.8 can be relaxed by replacing stable arguments by stable covers:

Theorem 4.3 *An assumption* δ *is conditionally entailed in a context* T *if and only if* δ *belongs to a stable cover in* T.

Theorem 4.3 provides a sound and *complete* syntactic characterization of the conditions under which an assumption can be legitimately adopted in a given context. While stable arguments are mostly adequate for capturing *linear* arguments,

stable covers provide the ability to deal adequately with *disjunctions* as example 4.7 illustrates.

Theorem 4.3, however, does not provide a characterization of the conditions under which *arbitrary* propositions are conditionally entailed by a given context. This extension, however, is straightforward. Let us say that a proposition p is *supported by a cover* C in a context $T = \langle E, K \rangle$ when every argument Δ in C is an argument *for* p, i.e. $E, \Delta \vdash_K p$. We have then the main result of this chapter:

Theorem 4.4 (Main) *A proposition p is conditionally entailed in a context T if and only if p is supported by a stable cover in T.*

We have thus arrived to a complete syntactic characterization of prioritized entailment in terms of admissible priority orderings. Once such a space of priority orderings is established, theorem 4.4 permits us to either analyze the truth conditions in the preferred classes of models of the contexts of interest, or determine the relevant stable arguments and covers and the propositions they support.

An undesirable feature of both approaches, though, is that they presume that we have identified the set of admissible priority orderings, and therefore, that we can check whether relations of the form $\Delta' \prec \Delta$ are necessarily satisfied. In many of the examples discussed in this chapter we have shown, however, that this is not a trivial task. Many times the constraints on admissible priority orderings imposed by a particular background contexts have a disjunctive form, and testing whether relations of the form $\Delta' \prec \Delta$ hold requires a good deal of work.

Fortunately, however, it is possible to replace such a test of by a corresponding *syntactic* test on K. For that purpose, we need to recall the definition of assumption dominance introduced in section 2.5 and reproduced below:

Definition 4.17 *A set Δ of assumptions* dominates *a set Δ' in a background context K, iff every assumption δ in Δ directly dominates $\Delta + \Delta'$.*

Recall that an assumption δ directly dominates arguments Δ in conflict with a default $p \rightarrow \delta$ in K (i.e., $p, \Delta \vdash_K \neg\delta$ and $p \not\vdash_K \neg\Delta$). We have appealed to this notion before for delimiting the space of priority orderings which are admissible with a given background context. It is not surprising thus, that due to the asymmetry and transitivity of priority orderings, the following syntactic characterization of the conditions under which a relation $\Delta' \prec \Delta$ must hold can be formulated:

Theorem 4.5 (Dominance) *For set of assumptions* Δ *and* Δ'*, the relation* $\Delta' \prec \Delta$ *holds in every priority ordering '*$\prec$*' admissible with a consistent background* $K = \langle L, D \rangle$ *if and only if* Δ *is part of a set that* dominates Δ' *in* K.

Theorems 4.5 and 4.3 together permit us to determine whether a given proposition is conditionally entailed by a consistent default theory by purely syntactic means. For that, we only need to look for stable covers and the corresponding dominance relations. Furthermore, we are now in a position to show that the irrelevance rule in the system **P** discussed in chapter 2, is indeed sound with respect to conditional entailment. Such a rule was introduced as an extension of the core and was responsible for drawing assumptions about independence. These assumptions amount to the following condition:

Theorem 4.6 (Irrelevance) *An assumption* δ *is conditionally entailed in a context* $T = \langle K, E \rangle$ *if for every argument* Δ' *against* δ*, there is a set* Δ*,* $\delta \in \Delta$*, that dominates* Δ' *in* K.

Given the soundness of the core for finite propositional languages, the following result is an straightforward consequence of the soundness of the irrelevance rule:

Theorem 4.7 *For finite propositional languages,* all *the rules of* **P** *are* sound *rules of conditional entailment.*

Such rules, however, are not *complete.* The irrelevance rule as defined in chapter 2 is only an approximation of the complete assertability conditions summarized in theorem 4.3. Indeed, examples such as 4.7, which require the notion stable covers, are beyond the power of rules 1–6.

4.3 Related Work

Conditional entailment is a refinement of an extension of the core recently advanced by Pearl [1989b], which is also related to a proposal due to Lehmann [1989]. Both accounts deal with finite propositional languages, where the core is sound and complete with respect to all forms of entailment analyzed in chapter 3, and both can be understood as suitable extensions of l-entailment. We will focus here on

Pearl's proposal only. If we recall from chapter 3, a layered-world structure is a pair $\langle \mathcal{W}, \kappa \rangle$, where $\mathcal{W}$ is a set of worlds, and κ is a world ranking function. A proposition p is l-entailed by a theory $T = \langle K, E \rangle$, when p is true in the preferred worlds of T in every layered-world structure admissible with K. Pearl extends l-entailment by considering only a *single* admissible layered world structure $\langle \mathcal{W}_{\mathcal{L}}, \kappa_{\min} \rangle$, where $\mathcal{W}_{\mathcal{L}}$ stands for the set of all possible worlds, and $\kappa_{\min}$ is a function which assign worlds a *minimal* rank.[14] For instance, for a language containing three propositional letters $\mathtt{p}$, $\mathtt{q}$ and $\mathtt{r}$, the minimal world structure $\langle \mathcal{W}_{\mathcal{L}}, \kappa_{\min} \rangle$ admissible with a background context K containing a single default $\mathtt{p} \rightarrow \mathtt{q}$, will be such that every world W that does not falsify $\mathtt{p} \rightarrow \mathtt{q}$ will have a rank $\kappa_{\min}(W) = 0$, and every world W' that falsify it will have a rank $\kappa_{\min}(W') = 1$. In particular, the lowest ranked world that satisfies both $\mathtt{p}$ and $\mathtt{r}$ will also satisfy the proposition $\mathtt{q}$, correctly legitimizing the conclusion $\mathtt{q}$ from $\mathtt{p}$ and $\mathtt{r}$; a conclusion which escapes l-entailment.

However, other inferences common to non-monotonic logics escape both accounts. For instance, given two defaults $\mathtt{p} \rightarrow \mathtt{q}$ and $\mathtt{p} \rightarrow \neg\mathtt{r}$ both accounts fail to authorize the conclusion $\mathtt{q}$ given both $\mathtt{p}$ and $\mathtt{r}$. The reason is that, in the resulting world ranking, the violation of one default "costs" as much as the violation of many defaults of equal rank. Similarly, while two defaults $\mathtt{p} \wedge \mathtt{s} \rightarrow \mathtt{q}$ and $\mathtt{r} \rightarrow \neg\mathtt{q}$, render the status of $\mathtt{q}$ ambiguous in the presence of $\mathtt{p}$, $\mathtt{s}$ and $\mathtt{r}$, such an ambiguity is resolved in favor of $\mathtt{q}$, when an additional default $\mathtt{p} \rightarrow \neg\mathtt{q}$ supporting its negation is added. Conditional entailment avoids these anomalies by ignoring defaults (assumptions) commonly violated by the pair of models being compared, and by considering *multiple* partial orders, as opposed to a *single* world ranking.

Outside the conditional camp, conditional entailment is closest to prioritized circumscription. Prioritized circumscription is a refinement of parallel circumscription, originally proposed by McCarthy [1986], and later developed by Lifschitz [1985, 1988a]. Roughly, the effect of prioritized circumscription is to induce a preference for models that assign smaller extensions to predicates with higher priorities. The only difference between conditional entailment and prioritized circumscription in the propositional case, is the source of such priorities: while prioritized circumscription relies on the user, conditional entailment automatically extracts the priorities from the knowledge base.

Two other technical differences arise, however, in the first-order case. First, the priorities in prioritized circumscription are on *predicates* as opposed to *atoms*. Such a difference often translates into a different behavior. For instance, in the

[14] Pearl call the resulting entailment relation 1-entailment which should not be confused with the entailment relation associated with layered world structures which call l-entailment.

"birds fly, penguins don't" example, the conclusion $\neg\mathtt{flies}(\mathtt{tim})$ is conditionally entailed by $\mathtt{penguin}(\mathtt{tim})$ by virtue of the priority of the assumption $\delta_2(\mathtt{tim})$ ('if Tim is a penguin, Tim does not fly') over the assumption $\delta_1(\mathtt{tim})$ ('if Tim is a bird, Tim flies'). The same behavior would normally be accommodated in prioritized circumscription by *assigning* a higher priority to the *predicate* $\mathtt{ab}_2 = \neg\delta_2$ than to the *predicate* $\mathtt{ab}_1 = \neg\delta_1$. However, such an encoding produces an unexpected behavior which does not arise in conditional entailment: given that either Tim is a flying penguin or that Tweety is a non-flying bird, for instance, prioritized circumscription is forced to conclude that Tweety is a non-flying bird.[15]

The second technical difference between conditional entailment and prioritized circumscription is the notion of *minimality* employed. In conditional entailment a model M of T is minimal iff it is has a minimal gap $\Delta[M]$; namely, if there is no model M' of T with violates a set of assumptions $\Delta[M']$ properly included in $\Delta[M]$. In particular, since assumptions are *ground literals*, M will be a minimal model of a theory $T = \{\exists\mathtt{x}.\,\neg\delta_1(\mathtt{x})\}$ iff M satisfies every assumption in the language. Every such model will thus presume the existence of one or many *unnamed* individuals in their respective domains which belong to the extension of the predicate δ_1. So while the formula $\delta_1(\mathtt{a})$ will hold in all minimal models of T, the formula $\exists\mathtt{x}.\,\forall\mathtt{y}.\,\neg\delta_1(\mathtt{y}) \Rightarrow \mathtt{y} = \mathtt{x}$ will not. The opposite is true in circumscription, where the minimality of a model is understood *semantically*, rather than *syntactically* (see for example, [Lifschitz, 1985]). In such a case no attention is paid to literals, but to individuals in the domains of the interpretations.

Which notion of minimality is preferred? The consensus is overwhelming in favor of a minimality notion understood in a semantic sense. However, if we reject the idea that such a choice carries a particular epistemological significance and that model theory and meaning are the same thing, the choice remains simply a selection of the most convenient device for formalizing default inference. So, which minimality criterion is more convenient for such a task? My view is in favor of a *syntactic* minimality criterion, as it permits us to reason about equality. Namely, given the negation of an assumption $\delta_1(\mathtt{a})$, we can still infer that an assumption $\delta_1(\mathtt{b})$ holds. A semantic minimality criterion, on the other hand, would require the inequality between $\mathtt{a}$ and $\mathtt{b}$ to be stated explicitly, precluding the possibility of $\mathtt{a}$ and $\mathtt{b}$ denoting the same thing. Similarly, given a default "birds fly," conditional entailment, unlike circumscription, is not bound to conclude that *all* birds fly. Indeed, the treatment of equality and universals in conditional entailment is closer

[15]Vladimir Lifschitz has recently brought to my attention the circumscriptive framework elaborated in [Lifschitz, 1988a] which permits a finer grained specification of priorities which avoids these anomalies.

to Reiter's [1980] default logic than to circumscription.

There are, however, severe limitations that arise from the focus on *literals* as opposed to *individuals*. Sometimes, we do want to assume that a property about all *individuals*. For instance, when reasoning about time we may want to assume that a clean block will remain clean unless a relevant change takes place. However, a default *schema* such as $\mathtt{on}(x, y, t) \rightarrow \mathtt{on}(x, y, t+1)$ would *not* authorize us to infer $\exists \mathtt{x}, \mathtt{y}\ \mathtt{on}(\mathtt{x}, \mathtt{y}, t+1)$ from $\exists \mathtt{x}, \mathtt{y}\ \mathtt{on}(\mathtt{x}, \mathtt{y}, t)$.[16] In that case, we do want to minimize the *extension* of predicate $\mathtt{ab_i} = \neg\delta_\mathtt{i}$ associated with the persistence of the on relation. Does that mean that we are forced back into a semantic notion of minimality? Not necessarily. It is possible to retain a syntactic minimality criterion and still be able to accommodate these forms of *closed world reasoning*.

Let us say that we want a predicate δ_i to be *closed* when we want the *extension* of δ_i to be *maximal* (i.e. the *complement of* δ_i to be *minimal*). Furthermore, let $\delta_i[M]$ stand for the set of tuples of ground terms t in the language such that $\delta_i(t) \in \Delta[M]$. Then, in order to *close a predicate* δ_i, it is sufficient to prune all those models M of the theory T of interest which fail to satisfy the condition $\forall x.\, \delta_i(x) \Rightarrow x \in \delta_i[M]$. If we say that a model of T is a model of the *closure* of T when all these closure conditions are satisfied, no empty gap model of a theory $T = \{\exists \mathtt{x}.\, \neg\delta_\mathtt{i}(\mathtt{x})\}$, for instance, would remain a model of the closure of T. Similarly, if the preferred models of T are selected among the models of the closure of T, a theory $T = \{\neg\delta_\mathtt{i}(\mathtt{a})\}$ will certainly conditionally entail the sentence $\forall \mathtt{x}.\, \mathtt{x} \neq \mathtt{a} \Rightarrow \delta_\mathtt{i}(\mathtt{x})$ very much as the circumscription of the complement of $\delta_\mathtt{i}$ will. So, it is possible to retain the appealing treatment of equality and universals of Reiter's default logic, and yet accommodate the form of closed world reasoning characteristic of circumscription.

In light of the relation between the model theory of prioritized circumscription and conditional entailment, it is not surprising to find that their respective proof-theories are related as well. An elegant proof-theory for prioritized circumscription was recently developed by Baker and Ginsberg [1989]. Baker and Ginsberg address the case in which predicates are linearly ordered; namely, circumscribed predicates are drawn from sets $P_1, P_2, \ldots, P_n$ such that the priority of a predicate in a set P_i is higher than the priority of a predicate in a set P_j, for $1 \leq j < i \leq n$. While differing in technical detail, the proof-theory they present has the same *dialectical* flavor as the proof-theory developed in section 4.3, which as they note, is closely related to approaches to defeasible inference based on the evaluation of arguments (e.g. [Loui, 1987a, Pollock, 1987]). The with Baker and Ginsberg are

[16] The example is from [McCarthy, 1980].

mainly in the treatment of disjunctions, which in our case, are pushed completely into what we called *covers*. Likewise, due to the nature of the constraints on admissible conditional priority orderings, we are forced to consider *sets* of non necessarily linear priority orderings. In this regard, the results in section 4.3 may prove relevant to prioritized circumscription, as they relax some of the assumptions on which the proof-theory of Baker and Ginsberg is based.

Chapter 5

The Causal Dimension: Evidence vs. Explanation

Conditional entailment extends preferential entailment and ϵ-entailment with many desirable features such as independence assumptions and default contraposition. It also accepts an intuitive proof-theory, akin to argument-based systems, in which arguments supporting incompatible propositions compete and those arguments based on assumptions of higher priority win. Nonetheless, while able to capture the intended behavior in a variety of contexts, many simple scenarios remain beyond its reach. *Unintended* models often slip into the set of *preferred* models, rendering a behavior weaker than expected.

In this chapter we argue that many of these problems arise from the *causal* nature of common defaults, which the conditional interpretation presented ignores. We claim that such a causal dimension of defaults is often the critical feature that distinguishes *intended* from *unintended* classes of models, and that it manifests itself in the fact that intended classes of models usually *cohere*. Namely, intended classes explain *why* certain expectations fail, like the failure of a car to get started due to a dead battery, or the change in position of a block due to a moving action.

Here we formalize these intuitions along two dimensions. On the one hand, we extend the language of default theories with a "causal" operator which allows the user to indicate when an exception is "explained." On the other, we use this operator to determine an ordering on classes of models which permits us to select among the conditional classes of a given theory, the most *coherent* ones.

Section 5.1 illustrates the limitations of conditional entailment and introduces the intuitions underlying the proposed refinement. Section 5.2 elaborates the language and interpretation of *causal* default theories. Finally, section 5.3 illustrates how tasks such as inheritance reasoning, reasoning about change, abductive reasoning, and reasoning about general logic programs, can all be expressed in the resulting framework.

5.1 Limitations of Conditional Entailment

A natural example in which to illustrate the limitations of conditional entailment is the now famous "Yale shooting problem" [Hanks and McDermott, 1986]. The problem describes a simple scenario where a gun loaded at a certain time (or situation) is shot at a person, Fred, at some later time. Hanks and McDermott devised the scenario and showed that a natural encoding in well-known non-monotonic formalisms — McCarthy's [1980, 1986] circumscription, Reiter's [1980] default logic and McDermott and Doyle [1980] non-monotonic logic — produced a behavior weaker than expected; namely, a behavior in which the fate of Fred remains undecided.

The Yale shooting problem spurred a large number of replies, ranging from special forms of circumscription, to alternative frameworks and encodings of the original problem. Some of of these replies have been reviewed by Hanks and McDermott themselves in [Hanks and McDermott, 1987]. We share the essence of Hanks and McDermott's argument that few of these proposals answer the problem in its full generality. Moreover, unlike Hanks and McDermott, we do not think that the temporal nature of the problem plays a distinctive role among the relevant issues. For that reason, we present a bare-bones propositional description of the Yale shooting scenario in which the same difficulties pointed out by Hanks and McDermott arise.

Example 5.1 (Yale Shooting Problem) The problem states that there is a gun loaded at time t, shot at a later time t' at a person alive at t. The question is to predict the fate of the person after the shooting. We encode the relevant relations in a background context $K = (L, D)$ with sentences (fig. 5.1):

$$\texttt{loaded} \wedge \delta_1 \Rightarrow \texttt{loaded}'$$
$$\texttt{alive}' \wedge \delta_2 \Rightarrow \texttt{alive}''$$
$$\texttt{shoot}' \wedge \texttt{loaded}' \Rightarrow \neg\texttt{alive}''$$

$\texttt{shoot}' \wedge \texttt{loaded}' \Rightarrow \neg\delta_2$

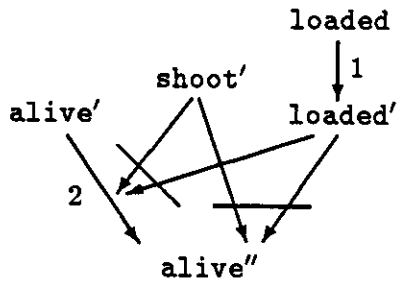


Figure 5.1: The "essential" Yale shooting problem

These sentences can be read as stating that the fluents `loaded` and `alive` tend to persist, and that a shooting with a loaded gun "clips" the fluent `alive`, switching its sign. Furthermore, we assume that K includes defaults of the form $p \rightarrow \delta_i$, for each of the sentences $p \wedge \delta_i \Rightarrow q$ above.

The target context $T = \langle K, E \rangle$ is defined by the evidence $E = \{\texttt{loaded}, \texttt{alive}', \texttt{shoot}'\}$. In such a context, two classes of minimal models arise: the expected class $\mathcal{C}_{\{2\}}$ in which δ_2 is the only violated assumption (i.e. `loaded` persists and `alive` does not), and the spurious class $\mathcal{C}_{\{1\}}$ in which δ_1 is the only violated assumption (i.e. `alive` persists and `loaded` does not). The former class would be preferred to the latter if the assumption δ_1 — the persistence of `loaded` — had priority higher than the assumption δ_2 — the persistence of `alive`. However, the background context K does not impose such a restriction, and as a result, like in the frameworks analyzed by Hanks and McDermott, neither $\neg\texttt{alive}''$ nor $\texttt{loaded}'$ are conditionally entailed by T.

Note that it is not unreasonable for the assumptions δ_1 and δ_2 to be unordered; after all, their status is normally independent: nor does the persistence of `loaded` normally presume the clipping of `alive`, nor does the persistence of `alive` normally presume the clipping of `loaded`. It is rather in the particular context in which a shooting aimed at a particular person takes place, in which the status of one assumption become relevant to the status of the other. Indeed, it is only in such context that it makes sense to regard one assumption as preferred to the other.

It is thus not surprising to find that such a behavior cannot be captured within the framework of conditional entailment in a natural way. Indeed, while the space

of priority orderings admissible with a given default theory $T = \langle K, E \rangle$ *depends* on the background K, it is *independent* of the particular body of evidence E considered. In particular, since in the absence of a `shooting` the fluents `loaded` and `alive` are unrelated, they remain unordered even when the evidence indicates that a `shooting` has taken place.

It is thus clear, that the particular body of evidence at hand needs also be taken into account for capturing the intuitive preferences among assumptions. As we will see, the need for such *dynamic* preferences also occurs in inheritance hierarchies, general logic programs, and reasoning to the best explanation. The following example is indeed a version of the Yale shooting problem which does not involve temporal persistences.

Example 5.2 Consider the following scenario. We get up in the morning and want to drive to work. However, we go to the car and notice that we have left the lights on for the whole night. At that point we want to assess the chances that the car is going to get started upon turning the ignition key. First, we assume that, normally, when the ignition key is turned the car engine starts. However, the car is likely not to start if the battery is dead. Moreover, the battery is likely to be dead after having left the lights on during the whole night. Thus K contains the following defaults (fig. 5.2):

$$\texttt{turn\_key} \rightarrow_1 \texttt{starts}$$
$$\texttt{turn\_key} \wedge \texttt{battery\_dead} \rightarrow_2 \neg\texttt{starts}$$
$$\texttt{lights\_were\_on} \rightarrow_3 \texttt{battery\_dead}$$

Namely, following the convention advanced in section 2.2, K contains a sentence $p \wedge \delta_i \Rightarrow q$ and a default $p \rightarrow \delta_i$ for each expression $p \rightarrow_i q$.

The context $T = \langle K, E \rangle$ with $E = \{\texttt{lights\_were\_on}, \texttt{turn\_key}\}$, gives rise to three minimal classes: $C_{\{1\}}$, $C_{\{2\}}$, and $C_{\{3\}}$, where $C_{\{i\}}$ stands for the class of models which violate the assumption δ_i only. Models in the class $C_{\{1\}}$ are preferred to models in the class $C_{\{2\}}$, as the assumption δ_2 is preferred to the assumption δ_1 in every priority order admissible with K. No such preference exists however between the assumptions δ_1 and δ_3, and thus, models in $C_{\{1\}}$ and $C_{\{3\}}$ remain equally preferred. As a result, conditional entailment sanctions the disjunction $\neg\texttt{starts} \vee \neg\texttt{battery\_dead}$, rather than the stronger, expected conclusion $\neg\texttt{starts}$. Note that, again, while unrelated in K, it seems reasonable to regard the assumptions δ_1 — about the expected state of the car after turning the ignition key — as preferred to the assumption δ_3 — about the expected state of

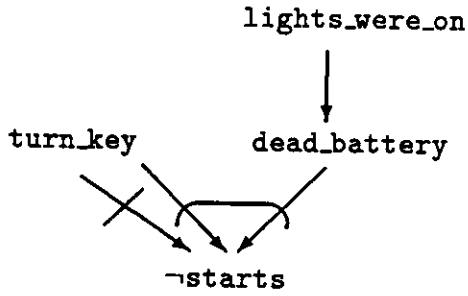


Figure 5.2: The battery problem

the battery after leaving the lights on. Indeed, while it is reasonable to *explain* the car not getting starting due to a dead battery, we cannot *explain* the battery not dying by predicting that the car will get started. Like in the Yale shooting problem, the *logical symmetry* between a pair of conflicting assumptions hides a *causal asymmetry* which underlies the intended behavior.

Example 5.3 As a last example, consider a scenario in which Mary is organizing a party to which she plans to invite most of her friends, who, normally, are likely to attend. However, people are likely not attend a party which is attended by somebody they dislike. Tammy and Peter are Mary's friends and Tammy dislikes Peter.

We express this knowledge as a default theory $T = \langle K, E \rangle$, with an evidence set $E = \{\texttt{friend(t)}, \texttt{friend(p)}, \texttt{dlikes(t,p)}\}$ and a background context $K = \langle L, D \rangle$ given by the following expressions (fig. 5.3):

$$\texttt{friend}(x) \rightarrow_1 \texttt{invited}(x)$$
$$\texttt{invited}(x) \rightarrow_2 \texttt{attends}(x)$$
$$\exists \texttt{y}.\, [\texttt{dlikes}(x, y) \wedge \texttt{attends}(y)] \rightarrow_3 \neg\texttt{attends}(x)$$

In the context T, it is reasonable to expect that both Tammy and Peter are going to be invited, that Peter is going to attend, while leaving in doubt whether Tammy will attend, considering her dislike for Peter. However, none of these conclusions are conditionally entailed by T. Indeed, there are five minimal classes, each violating an assumption from the set $\delta_1(\texttt{t}), \delta_1(\texttt{p}), \delta_2(\texttt{t}), \delta_2(\texttt{p}), \delta_3(\texttt{t})\}$, *all equally preferred*. Again, the intended classes are precisely those in which the violated

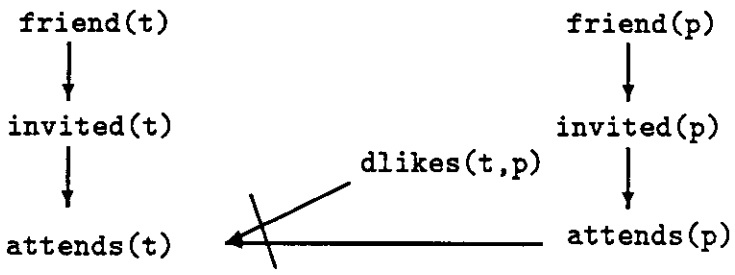


Figure 5.3: The party problem

assumptions accept explanations in terms of the available information: e.g. Tammy did not go *because* Peter was attending and Tammy does not like Peter; or Tammy did go in spite of Peter, because she usually attends Mary's parties. The intended classes, on the other hand, include models in which Peter is invited but does not go, or where he is not even invited.[1]

All these examples illustrate contexts which generate spurious preferred classes and, therefore, weak conclusions. We claim that that characteristic feature that distinguishes intended form unintended models is *coherence*: while 'exceptions' are explained in the intended classes, they are *not* explained in the unintended ones. The refinement of conditional entailment to be developed in the next section is based on an extension of the language of default theories which will permit the user to express when an exception is explained. For a (plain) formula α, there will be be a new formula $C\alpha$ in the language, which will be read as stating that α is 'explained.' The coherence of a class of models will then be determined by considering the truth of the literals $C\neg\delta$ for the assumptions δ violated in the class. These considerations will allow us to prune the set of conditional models associated with a given context, and to strengthen the resulting defeasible entailment relation. We will later show that the resulting framework goes well beyond the examples considered in this section, and provides some useful insights into the semantic issues surrounding the use of negation in logic programming (section 5.3.3), as well as on issues related to abductive reasoning (section 5.3.4).

[1]These expectation failures could be explained in principle by postulating appropriate hypotheses. For example, we could explain that Peter did not go in spite of being invited, in order not to affect Tammy's chances of attending. However, we only consider here explanations which do not need hypotheses unsupported by the available evidence. Such explanations will be considered in section 5.3.4 in the context of abductive reasoning.

5.2 Causal Theories

5.2.1 Language

A causal theory is a default theory whose underlying language $\mathcal{L}$ has been extended with the causal operator 'C.' We refer to the resulting language as $\mathcal{CL}$. We call the formulas in $\mathcal{L}$ *plain* formulas, while those in $\mathcal{CL}$ but not in $\mathcal{L}$ as *causal* formulas. The language $\mathcal{CL}$ is closed under all the standard classical connectives. On the other hand, we do not allow embedded causal operators; thus if γ is a causal formula, $C\gamma$ will not be in $\mathcal{CL}$.

For example the *causal encoding* of the version of the Yale shooting problem considered above will contain the following rules:

$$\texttt{loaded} \wedge \delta_1 \Rightarrow \texttt{loaded}'$$
$$\texttt{alive}' \wedge \delta_2 \Rightarrow \texttt{alive}''$$
$$\texttt{shoot}' \wedge \texttt{loaded}' \Rightarrow C\neg\texttt{alive}''$$
$$\texttt{shoot}' \wedge \texttt{loaded}' \Rightarrow C\neg\delta_2$$

Namely, rules of the form 'if p then q' — such as the shooting rule — which have a causal or explanatory character, are encoded as *causal* rules of the form 'if p then Cq' which can be read as stating that if p is true, then q is explained.

We will assume that the operator 'C' complies with certain minimal restrictions which happen to correspond to the postulates of the system T in modal logic [Hughes and Cresswell, 1968]:

[C1] $C\alpha \Rightarrow \alpha$
[C2] $C(\alpha \Rightarrow \beta) \Rightarrow (C\alpha \Rightarrow C\beta)$
[C3] If $\vdash_K \alpha$ then $C\alpha$

[C1] forces every explained proposition to be true, [C3] renders every expression that logically follows from the background as explained, while [C2] guarantees that the set of explained proposition is closed under deduction.

The operator 'C' will be used to induce an order on *classes* of models, and indirectly, to determine a set of *causally preferred* models.

5.2.2 Semantics: Causal Entailment

Let us first recall that for an interpretation M, $\Delta[M]$ refers to the assumptions violated in M. Every assumption δ not in $\Delta[M]$ is thus an assumption that holds in M. A class $\mathcal{C}$ of models with an associated gap Δ stands for the collection of models M of T such that $\Delta[M] \subseteq \Delta$. The class $\mathcal{C}$ is *minimal* iff for every model M in $\mathcal{C}$, $\Delta[M] = \Delta$. Likewise, a proposition p holds in $\mathcal{C}$ iff it holds in every model in $\mathcal{C}$. Proof-theoretically, this condition is equivalent to the existence of a classical derivation of p from the sentences in T and assumptions not in the gap of $\mathcal{C}$. Let us also recall that a *conditional model (class)* in a context T, refers to a model (class) preferred in some admissible prioritized structure. Similarly, we refer to an admissible priority ordering as a *conditional* ordering.

As usual, we call the complement of an assumption δ an 'exception' or an 'abnormality.' Furthermore, we will say that an exception $\neg\delta$ is *explained in a class*, when the causal literal $C\neg\delta$ holds in the class. From the remarks above, a sufficient and necessary condition for an 'exception' $\neg\delta$ to be explained in a class $\mathcal{C}$ in a context T, is thus the existence of a set of assumptions not in gap of $\mathcal{C}$ Δ, that together with T imply the literal $C\neg\delta$. Thus all explanations in a class are grounded in the available evidence and the assumptions validated by the class.

When the context T is understood, we will use the notation $\Delta^e[\mathcal{C}]$ to refer to the set of assumptions δ whose negations are explained in $\mathcal{C}$. We call such a set the *explained gap* of $\mathcal{C}$, an distinguish it from the *unexplained* gap of $\mathcal{C}$, given by the collection of assumptions in the gap of $\mathcal{C}$ but not in its explained gap. The unexplained gap of a class is thus a measure of its *incoherence*. In particular, when a class possesses an empty unexplained gap, the assumptions in the class fit perfectly well. We call such classes *perfectly coherent* classes. Intuitively, no class can be 'better' than a perfectly coherent class. The following *causal* preference relation on classes formalizes this intuition:

Definition 5.1 *Let $\mathcal{C}$ and $\mathcal{C}'$ be two classes in a context T. The class $\mathcal{C}$ is* as causally preferred as *the class $\mathcal{C}'$ iff $\Delta[\mathcal{C}] - \Delta^e[\mathcal{C}] \subseteq \Delta[\mathcal{C}']$. The class $\mathcal{C}$ is* causally preferred *to $\mathcal{C}'$ iff $\mathcal{C}$ is as causally preferred as $\mathcal{C}'$ but $\mathcal{C}'$ is* not *as causally preferred as $\mathcal{C}$.*

In words, a class $\mathcal{C}$ is causally preferred to $\mathcal{C}'$ when every exception in $\mathcal{C}$ but not $\mathcal{C}'$ has an explanation, but not vice versa. We will also say that $\mathcal{C}$ is a *(causally) preferred* class in the context T, iff there is no other class in T causally preferred to

$\mathcal{C}$. We call $\mathcal{C}$ a *causal class* of T, and refer to the models it contains as *causal models* of T. It is simple to show that causal models and classes are minimal; namely, among the models and classes of T they possess a minimal gap. Furthermore, even though the causal preference relation on classes *is not necessarily transitive*, the causal classes of T can be determined by considering the minimal classes of T only.

In analogy to conditional entailment, *causal entailment* is defined as follows:

Definition 5.2 *A theory T* causally entails (cs-entails) *a proposition p iff p holds in all the causally preferred classes of T.*

We will illustrate these definitions with some examples. Later on we will integrate causal entailment and conditional entailment into a single entailment relation which takes into account both causal and conditional aspects of defaults.

Example 5.4 Let us consider first a theory T given by the single causal sentence $\delta \Rightarrow \mathrm{C}\neg\delta'$, where δ and δ' are different assumptions. Such a context admits two minimal classes: a class $\mathcal{C}$ with an associated gap $\Delta[\mathcal{C}] = \{\delta'\}$, and a class $\mathcal{C}'$ with an associated gap $\Delta[\mathcal{C}'] = \{\delta\}$. Both classes represent all the minimal classes of T, as there is no model of T that satisfies both δ and δ', and the restrictions [C1]–[C3]. The class $\mathcal{C}$ is committed to the assumption δ while the class $\mathcal{C}'$ is committed to the assumption δ'. Furthermore, since $T, \delta \vdash \mathrm{C}\neg\delta'$ holds, it follows that the exception $\neg\delta'$ is explained in $\mathcal{C}$. On the other hand, the exception $\neg\delta$ is *not* explained in $\mathcal{C}'$, as there is no assumption set validated by $\mathcal{C}'$ which supports the literal $\mathrm{C}\neg\delta$. It follows then, that the class $\mathcal{C}$ is causally preferred to $\mathcal{C}'$, as $\Delta[\mathcal{C}] - \Delta^{\mathrm{c}}[\mathcal{C}] = \emptyset \subseteq \Delta[\mathcal{C}']$, but $\Delta[\mathcal{C}'] - \Delta^{\mathrm{c}}[\mathcal{C}'] = \{\delta\} \not\subseteq \Delta[\mathcal{C}] = \{\delta'\}$. Furthermore, since $\mathcal{C}$ and $\mathcal{C}'$ are the only minimal classes of T, $\mathcal{C}$ remains as the single causally preferred class of T, and the propositions δ and $\neg\delta'$ are causally entailed (cs-entailed) by T.

Note the asymmetry established by the causal preferences on the theory T; while the assumptions δ and δ' are incompatible, δ is cs-entailed but δ' is not.

The next example illustrates the behavior of the causal formulation of the Yale shooting scenario described above.

Example 5.5 The causal formulation consists of the following sentences:

$$\texttt{loaded} \wedge \delta_1 \Rightarrow \texttt{loaded}'$$

$$\texttt{alive}' \wedge \delta_2 \Rightarrow \texttt{alive}''$$
$$\texttt{shoot}' \wedge \texttt{loaded}' \Rightarrow \mathrm{C}\neg\texttt{alive}''$$
$$\texttt{shoot}' \wedge \texttt{loaded}' \Rightarrow \mathrm{C}\neg\delta_2$$

The difference with the formulation in example 5.1, is that the causal character of the shooting rule has been made explicit . As usual, we also have defaults $p_i \rightarrow \delta_i$ for each of the sentences of the form $p_i \wedge \delta_i \Rightarrow q_i$ in K.

The target context $T = \langle K, E\rangle$ with $E = \{\texttt{loaded}, \texttt{alive}', \texttt{shoot}'\}$, gives rise to two minimal classes: $\mathcal{C}_{\{1\}}$, where the assumption δ_2 about the persistence of alive holds and the assumption δ_1 is violated; and $\mathcal{C}_{\{2\}}$, where the assumption δ_1 about the persistence of loaded holds, and the assumption δ_2 is violated. These two classes, however, are no longer symmetrical. Indeed, while the assumption δ_1 *explains* the violation of δ_2, namely, $E, \delta_1 \models_K \mathrm{C}\neg\delta_2$, no set of assumptions valid in $\mathcal{C}_{\{1\}}$ supports an explanation for $\neg\delta_1$. Thus, we obtain $\Delta[\mathcal{C}_{\{2\}}] = \Delta^c[\mathcal{C}_{\{2\}}] = \{\delta_2\}$, while $\Delta[\mathcal{C}_{\{1\}}] = \{\delta_1\}$ and $\Delta^c[\mathcal{C}_{\{1\}}] = \emptyset$. As a result, the minimal class $\mathcal{C}_{\{2\}}$ is perfectly coherent and is causally preferred to the only other minimal class $\mathcal{C}_{\{1\}}$. It follows then that the propositions that hold in $\mathcal{C}_{\{2\}}$ are causally entailed by T and, in particular, the propositions loaded′ and $\neg$alive″.[2] [3]

[2]Other solutions to the Yale shooting problem which rest on the same intuition of minimizing, or even banishing unexplained abnormality are Lifschitz's [1987], Haugh's [1987], and Morgenstern's and Stein's [1988].

[3]It should be mentioned that there are logically *consistent* causal theories, which lack *causal* models. This is a consequence of the fact that causal preferences on classes are not always transitive. Consider for instance a context T given by the rules $\delta_1 \Rightarrow \mathrm{C}\neg\delta_2$, $\delta_2 \Rightarrow \mathrm{C}\neg\delta_3$, and $\delta_3 \Rightarrow \mathrm{C}\neg\delta_1$. Such a context admits three minimal classes $\mathcal{C}_{\{1,2\}}$, $\mathcal{C}_{\{1,3\}}$, and $\mathcal{C}_{\{2,3\}}$, where $\mathcal{C}_{\{i,j\}}$ stands for a class with gap $\{\delta_i, \delta_j\}$. It is simple to show that the class $\mathcal{C}_{\{1,2\}}$ is causally preferred to $\mathcal{C}_{\{2,3\}}$, that $\mathcal{C}_{\{2,3\}}$ is causally preferred to $\mathcal{C}_{\{1,3\}}$, and that $\mathcal{C}_{\{1,3\}}$ is causally preferred to $\mathcal{C}_{\{1,2\}}$. These preferences, thus, establish a loop. As a result, T does not possess any causally preferred class, and hence, it entails any sentence in the language.

There is a simple refinement of the definition of causal entailment, originally proposed in [Geffner, 1989], which avoids these anomalies. Let $\mathcal{C}$ be a class in a context T, and let us say that a class $\mathcal{C}$ is *causally admissible* in T, iff every class $\mathcal{C}'$ in T which is causally preferred to $\mathcal{C}$ is such that $\Delta[\mathcal{C}'] \subset \Delta[\mathcal{C}]$. In the example above, the class $\mathcal{C}_{\{1,2,3\}}$ with gap $\{\delta_1, \delta_2, \delta_3\}$, is a causally admissible class, while none of the minimal classes $\mathcal{C}_{\{1,2\}}$, $\mathcal{C}_{\{1,3\}}$ and $\mathcal{C}_{\{2,3\}}$ is. Then a form of *cautious* causal entailment can be defined, where rather than focusing on the causally preferred classes of T, we only consider its *minimal admissible* classes. The resulting consequence relation is guaranteed to be well-behaved—namely, consistent whenever T is consistent— as every consistent T has at least one admissible class: the class $\mathcal{C}_{\Delta_{\mathcal{CL}}}$, whose gap $\Delta_{\mathcal{CL}}$ contains all the assumptions in the language.

5.2.3 Integrating Causal and Conditional Preferences

The notion of causal entailment above was motivated as a refinement of the notion of conditional entailment developed in the last chapter. We argued that in addition to the conditional dimension of default reasoning, there is an additional dimension of defaults related to causality and coherence, which pops up in problems such as those considered in section 5.1. Here we will present a simple scheme that integrates both causal and conditional considerations.[4]

Causal entailment (cs-entailment) is defined in terms of the causally preferred classes of a given theory T. These causally preferred classes are drawn from the space of all classes associated with T or, with identical results, from the space of *minimal* classes of T. The notion of conditional entailment (cd-entailment) developed in chapter 4 was introduced in an analogous way, selecting the *conditional* rather than the *causal* classes of T.

In order to refine the conditional entailment relation we will apply the causal ordering to refine the set of conditional classes of T. In other words, we will use the causal ordering to select, *among the conditional classes of* T, the causally preferred ones. Naturally, we refer to the resulting classes as the *causal conditional classes* of T and to the resulting entailment relation as *causal conditional entailment (csd-entailment)*. Thus we say that a context T csd-entails a proposition p when p holds in all the *causally preferred conditional classes* of T.

It is straightforward to show that T cd-entails p only if T csd-entails p. Not the other way around though, as the Yale shooting scenario reveals (example 5.5). Such an example, however, did not display interactions between causal and conditional preferences, as all the minimal classes were also conditional ones. These interactions are illustrated in the reformulation below.

Example 5.6 We consider the same formulation employed in example 5.6 above, except that we assume now that the effects of the shooting rule hold by default:

$$\texttt{loaded} \wedge \delta_1 \Rightarrow \texttt{loaded}'$$
$$\texttt{alive}' \wedge \delta_2 \Rightarrow \texttt{alive}''$$
$$\texttt{shoot}' \wedge \texttt{loaded}' \wedge \delta_3 \Rightarrow \mathrm{C}\neg\texttt{alive}''$$
$$\texttt{shoot}' \wedge \texttt{loaded}' \wedge \delta_3 \Rightarrow \mathrm{C}\neg\delta_2$$

Each of the first three rules $p_i \wedge \delta_i \Rightarrow q_i$ in K has also an associated default of the form $p_i \rightarrow \delta_i$.

[4]Other integration schemes are also possible. See section 6.2.

We consider again the context $T = \langle K, E \rangle$, with $E = \{\texttt{loaded}, \texttt{alive}', \texttt{shoot}'\}$. In order to determine which propositions are csd-entailed by T, we have to identify the conditional classes of T and select, among them, the causally preferred ones. Since conditionally and causally preferred classes are guaranteed to be minimal, it is natural to start with the minimal classes.

The context T gives rise to three minimal classes $\mathcal{C}_{\{1\}}$, $\mathcal{C}_{\{2\}}$, and $\mathcal{C}_{\{3\}}$, where each $\mathcal{C}_{\{i\}}$ stands for the class of models in which only the assumption δ_i is violated. However, $\mathcal{C}_{\{3\}}$ is not a conditional class; the class $\mathcal{C}_{\{2\}}$ is conditionally preferred to class $\mathcal{C}_{\{3\}}$ as the assumption δ_3 is conditionally preferred to the assumption δ_3. Namely, the assumption δ_3 directly dominates the assumption set $\{\delta_2\}$ as a result that $\texttt{shoot}', \texttt{loaded}', \delta_2 \mid\!\sim_K \neg\delta_3$ holds, but $\texttt{shoot}', \texttt{loaded}' \mid\!\sim_K \neg\delta_2$ does not. Thus, the causally preferred conditional classes of T are among $\mathcal{C}_{\{1\}}$ and $\mathcal{C}_{\{2\}}$. As illustrated in example 5.5 above, the class $\mathcal{C}_{\{2\}}$ is causally preferred to $\mathcal{C}_{\{2\}}$, and thus, T csd-entails every proposition that holds in $\mathcal{C}_{\{2\}}$.

Although the causal encoding of the scenarios we have so far considered has been straightforward, often some care is required. The next example illustrates some of the aspects that need to be taken into account for such a translation to yield the expected behavior. More general guidelines will be analyzed in the next section.

Example 5.7 Consider a causal formulation of the "battery" example discussed above. The background context is described by the sentences:

$$\texttt{turn\_key} \wedge \delta_1 \Rightarrow \mathrm{C}\texttt{starts}$$
$$\texttt{turn\_key} \wedge \texttt{battery\_dead} \wedge \delta_2 \Rightarrow \mathrm{C}\neg\texttt{starts}$$
$$\texttt{lights\_were\_on} \wedge \delta_3 \Rightarrow \mathrm{C}\texttt{battery\_dead}$$

together with defaults $p_i \rightarrow \delta_i$ for each sentence $p_i \wedge \delta_i \Rightarrow q_i$. The rules are in causal form, because we agree to regard their consequents as explained when their antecedents hold.

Let us consider a context $T = \langle K, E \rangle$, with $E = \{\texttt{lights\_were\_on}, \texttt{turn\_key}\}$. As before, T gives rise to three minimal classes $\mathcal{C}_{\{1\}}$, $\mathcal{C}_{\{2\}}$ and $\mathcal{C}_{\{3\}}$, where $\mathcal{C}_{\{i\}}$ stands for the class which violates the assumption δ_i, from which only $\mathcal{C}_{\{1\}}$ and $\mathcal{C}_{\{3\}}$ are conditionally preferred. Since $\texttt{dead\_battery}$ holds in $\mathcal{C}_{\{1\}}$ we would expect the failure of the car to get started ($\neg\delta_1$) to be explained in $\mathcal{C}_{\{1\}}$; however, the literal $\mathrm{C}\neg\delta_1$ does not hold in $\mathcal{C}_{\{1\}}$, the class $\mathcal{C}_{\{1\}}$ is not preferred to $\mathcal{C}_{\{3\}}$, and the expected conclusion $\neg\texttt{starts}$ is not sanctioned.

Note that the discrepancy between the expected behavior and the actual behavior does not refute the intuition that classes should be ordered in terms of coherence considerations. It rather exposes the fact that the causal interpretation of T fails to uncover relevant explanatory patterns. The problem in this case, is that while `dead_battery` is able to explain the failure of `starts`, it is unable to explain the failure of the assumption (δ_1) which *predicts* it.

The fix we are about to propose rests on a simple intuition: that the assumption δ_i associated with the encoding of a default $p \rightarrow_i q$ should be explainable from p and an *explanation of the negation of* q. Such an intuition demands that a default $p \rightarrow_i q$ be encoded not only in terms of the pair of expressions $p \wedge \delta_i \Rightarrow q$ (or, $p \wedge \delta_i \Rightarrow \mathrm{C}q$) and $p \rightarrow \delta_i$, *but also in terms of an additional causal sentence* $p \wedge \mathrm{C}\neg q \Rightarrow C\neg\delta_i$.[5]

In the context of the current example, this encoding requires the inclusion in K of the following three additional rules:[6]

$$\texttt{turn\_key} \wedge \mathrm{C}\neg\texttt{starts} \Rightarrow \mathrm{C}\neg\delta_1$$
$$\texttt{turn\_key} \wedge \texttt{battery\_dead} \wedge \mathrm{C}\texttt{starts} \Rightarrow \mathrm{C}\neg\delta_2$$
$$\texttt{lights\_were\_on} \wedge \mathrm{C}\neg\texttt{battery\_dead} \Rightarrow \mathrm{C}\neg\delta_3$$

The theory T' that results from such an encoding gives rise again to the same three minimal classes $\mathcal{C}_{\{1\}}$, $\mathcal{C}_{\{2\}}$ and $\mathcal{C}_{\{3\}}$, only two of which, $\mathcal{C}_{\{1\}}$ and $\mathcal{C}_{\{3\}}$ are conditionally preferred. However, now the literal $\mathrm{C}\neg\delta_1$ holds in every model in $\mathcal{C}_{\{1\}}$, and thus $\mathcal{C}_{\{1\}}$ is causally preferred to $\mathcal{C}_{\{3\}}$. As a result, $\mathcal{C}_{\{1\}}$ is the causally preferred conditional class of T', and the propositions $\neg$`starts` and `battery_dead` are legitimized.

The examples illustrate that the introduction of the causal modal operator C provides a more powerful representational language in which to express intuitive patterns of inference which escape the machinery of conditional entailment. The inconvenience, however, is that the commonsense interpretation of the operator C involve fuzzy notions of explanation and causality which make the encoding of knowledge in the form of causal theories somewhat arbitrary. In the example

[5] Gelfond [1989] and Konolige and Myers [1989] suggest related encodings of defaults in autoepistemic logic. We will discuss the relation between causal and autoepistemic theories in section 5.4.

[6] Note that we write C`starts`, for instance, rather than C$\neg\neg$`starts`. This is only a convenience. Both causal sentences possess the same models under the constraints on C.

above, for instance, we claimed that a default $p_i \rightarrow_i q_i$ should be expressed as a pair of sentences $p_i \wedge \delta_i \Rightarrow \mathrm{C}q_i$ and $p_i \wedge \mathrm{C}\neg q_i \Rightarrow \mathrm{C}\neg\delta_i$, together with a default $p_i \rightarrow \delta_i$. This raises the questions of whether this encoding, among the many possible, is a 'trick" contrived to make the proposed interpretation work, or to the contrary, it fits within more general guidelines for expressing defeasible knowledge in the form of causal theories. In the rest of this chapter we address this issue. We show that there are simple, *local* mappings (i.e. expression by expression) which permit expressing knowledge about a variety of domains in the form of causal theories, in such a way that the intended behavior is captured. We consider defeasible inheritance hierarchies, reasoning about change, general logic programs, and abductive reasoning.

5.3 Applications

5.3.1 Inheritance Hierarchies

Defeasible inheritance hierarchies are convenient devices for organizing knowledge about prototypical classes of individuals [Touretzky, 1986]. They take the form of directed graphs, where links connecting nodes x and y represent either that x is a member of the class y — if x represents an individual— or that members of the class x are normally members of the class y— if x stands for a class. For negated links connecting x to y, the same relations are to be understood in terms of x and the negation of y. Given an inheritance hierarchy Γ, the problem is to determine the properties which a given individual can be assumed to possess; a problem that amounts to determining which directed paths in the net encode legitimate inferences (see [Touretzky, 1986, Horty *et al.*, 1987, Geffner and Verma, 1989]).

Conditional entailment, while able to accommodate languages and patterns of inference that go beyond those captured by inheritance schemes, fails to provide an intuitive account of what they do capture. This is not surprising, though, as the links in the net represent more than conditional statements. There is also a causal component which needs to be taken into account in order to identify the inheritance relations embedded in a net. In the example below we illustrate where this component comes from and show how a simple mapping of inheritance hierarchies into causal theories solves the problem.

Example 5.8 Let us consider the inheritance hierarchy depicted in fig. 5.4. Such a

network can represent something like "USC students are students", "most students are unemployed," and while "unemployed are not rich," "USC students are." The natural encoding of such a network is in the form of a background context K with sentences:

$$\begin{aligned}
&\mathtt{A}(x) \wedge \delta_1(\mathtt{x}) \Rightarrow \mathtt{B}(x)\\
&\mathtt{B}(x) \wedge \delta_2(\mathtt{x}) \Rightarrow \mathtt{C}(x)\\
&\mathtt{C}(x) \wedge \delta_3(\mathtt{x}) \Rightarrow \neg\mathtt{D}(x)\\
&\mathtt{A}(x) \wedge \delta_4(\mathtt{x}) \Rightarrow \mathtt{D}(x)
\end{aligned}$$

and default schemas $p_i(x) \rightarrow \delta_i(x)$ for each such sentence $p_i(x) \wedge \delta_i(x) \Rightarrow q_i(x)$.

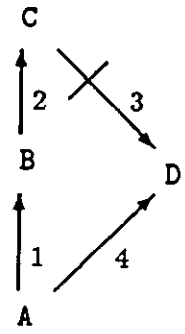


Figure 5.4: A simple network: A's are expected to be C's

In such a background context, we want to determine whether an instance, say a, of class A (e.g. "USC students"), is declared to be an instance of class C (e.g. "unemployed"). In the context $T = \langle K, E\rangle$, with $E = \{\mathtt{A}(\mathtt{a})\}$, the four assumptions $\delta_i(\mathtt{a})$, $i = 1, \ldots, 4$, are in conflict, giving thus rise to four minimal classes $\mathcal{C}_{\{i\}}$, $i = 1, \ldots, 4$, where $\mathcal{C}_{\{i\}}$ stands for the class of models in which the assumption $\delta_i(\mathtt{a})$ is violated. As shown in example 4.6, only two of these classes are conditionally preferred: $\mathcal{C}_{\{2\}}$ and $\mathcal{C}_{\{3\}}$. As a result, while the sentences B(a) and D(a) are conditionally entailed by A(a), the sentence C(a) unexpectedly is not.

The diagnosis of such misbehavior is not difficult to identify: conditional entailment fails to capture the intuition that while there is no justification for 'breaking' the argument path $\mathtt{B} \rightarrow \mathtt{C} \not\rightarrow \mathtt{D}$ between B and C, *there is* a justification for 'breaking' it between C and $\neg$D; the justification being the *conflicting* link $\mathtt{A} \rightarrow \mathtt{D}$. The

encoding of defaults as causal rules along the lines suggested in example 5.7, captures this intuition.

An inheritance network Γ will be mapped into a causal default theory $T = \langle K, E \rangle$ as follows. Defeasible links connecting a class p to a class q are tagged with a unique integer i, and are translated into a default schema $p(x) \rightarrow \delta_i(x)$ and a pair of causal rules $p(x) \wedge \delta_i(x) \Rightarrow \mathrm{C}q(x)$ and $p(x) \wedge \mathrm{C}\neg q(x) \Rightarrow \mathrm{C}\neg\delta_i(x)$. These rules will guarantee that 'exceptions' $\neg\delta_i(x)$ will be explained when conflicting directed paths to $\neg q$ are supported. Strict links connecting classes p and q, on the other hand, are mapped into universal sentences $p(x) \Rightarrow q(x)$. Finally, the evidence set E of T is simply the collection of atoms of the form $p(a)$, for individual nodes connected to class nodes in Γ. The problem of determining whether an individual a can be assumed to inherit a property q in Γ is thus mapped to the problem of determining whether the atom $q(a)$ is entailed by T; namely, whether $q(a)$ holds in all the causally preferred conditional classes of T.

Example 5.9 We consider now the causal encoding of the inheritance hierarchy Γ depicted in fig. 5.4. The background K of the causal theory T contains thus the causal rules:

$$\mathtt{A}(x) \wedge \delta_1(\mathtt{x}) \Rightarrow \mathrm{C}\mathtt{B}(x)$$
$$\mathtt{B}(x) \wedge \delta_2(\mathtt{x}) \Rightarrow \mathrm{C}\mathtt{C}(x)$$
$$\mathtt{C}(x) \wedge \delta_3(\mathtt{x}) \Rightarrow \mathrm{C}\neg\mathtt{D}(x)$$
$$\mathtt{A}(x) \wedge \delta_4(\mathtt{x}) \Rightarrow \mathrm{C}\mathtt{D}(x)$$

where for each sentence $p_i(x) \wedge \delta_i(x) \Rightarrow \mathrm{C}q_i(x)$, we add a sentence $p_i(x) \wedge \mathrm{C}\neg q_i(x) \Rightarrow \mathrm{C}\neg\delta_i(x)$ and a default schema $p_i(x) \rightarrow \delta_i(x)$. The context T contains in addition a body of evidence $E = \{\mathtt{A}(\mathtt{a})\}$.

The context T gives rise, again, to four minimal classes $\mathcal{C}_{\{i\}}$, $i = 1, \ldots, 4$, where $\mathcal{C}_{\{i\}}$ stands for class of models in which the assumption $\delta_i(\mathtt{a})$ is violated. Likewise, only the classes $\mathcal{C}_{\{2\}}$ and $\mathcal{C}_{\{3\}}$ remain conditionally preferred. However, this time, among the classes $\mathcal{C}_{\{2\}}$ and $\mathcal{C}_{\{3\}}$ only the latter is causally preferred. The reason is that we can explain $\neg\delta_3(\mathtt{a})$ (in terms of the assumptions $\delta_1(\mathtt{a})$, $\delta_2(\mathtt{a})$ and $\delta_4(\mathtt{a})$), but we cannot explain the exception $\neg\delta_2(\mathtt{a})$. As expected then, $\mathtt{A}(\mathtt{a})$ entails the propositions $\mathtt{B}(\mathtt{a})$, $\mathtt{D}(\mathtt{a})$, and $\mathtt{C}(\mathtt{a})$.

5.3.2 Reasoning about Change

Inheritance hierarchies are simple domains which involve a single type of defeasible statements. The situation is more complex when reasoning about change. Simple theories for reasoning about change need to represent the effects of actions, the conditions which can prevent actions from achieving their normal effects, and the tendency of certain aspects of the world (fluents) to remain stable in the absence of relevant changes (see [McDermott, 1982], for instance). We will refer to the first type of rules as *change* rules, to the second type as *cancellation* rules, and to the third type as *persistence* rules.

Change, cancellation and persistence rules can interact in various ways. The Yale shooting illustrates problems that result from an inadequate handling of the interactions between change and persistence rules. We showed, however, that it is possible to avoid these problems by expressing the Yale shooting problem as a causal theory. We now present general guidelines to locally map general theories for reasoning about change into causal theories.

Rules about change are encoded in a way similar to defeasible links in inheritance networks. Such encoding is uncommitted about the particular temporal notation used. For simplicity, we use a simple reified temporal language (see, [Shoham, 1987], for instance), sufficient to illustrate the relevant issues. Other notations could be used as well. The notation $p(x)_t$ below, where p is a predicate and t is a time point, is used as an abbreviation of the sentence Holds$(p(x), t)$, to read "fluent $p(x)$ holds at time t." We will also assume for simplicity a discrete time where t precedes $t+1$.

Formally, a rule about change of the form

$$precond(x)_t \wedge action(x)_t \rightarrow effect(x)_{t+1}$$

where $precond(x)$ stands for the (default) preconditions of $action(x)$, is mapped into a default causal rule; namely, a default schema

$$precond(x)_t \wedge action(x)_t \rightarrow \delta_i(x)_t$$

with a unique predicate δ_i, and sentences

$$precond(x)_t \wedge action(x)_t \wedge \delta_i(x)_t \Rightarrow C effect(x)_{t+1}$$

and

$$precond(x)_t \wedge action(x)_t \wedge \mathrm{C}\neg effect(x)_t \Rightarrow C\neg\delta_i(x)_t$$

Causal expressions of the form $\mathrm{C}p(x)_t$ can thus be understood as stating that $p(x)$ has been *caused to hold* at time t, or that $p(x)$ has been 'initiated' at time t, as in Kowalski and Sergot's [1986] event calculus.

The persistent of a fluent f (e.g. $\mathtt{on(a,b)}$), on the other hand, is expressed by a default of the form

$$f_t \rightarrow \delta(f)_t$$

where $\delta(f)_t$ stands for the atom $\delta(f,t)$, read "the *persistence* of f holds at time t," together with two sentences:

$$f_t \wedge \delta(f)_t \Rightarrow f_{t+1}$$

$$\mathrm{C}\neg f_{t+1} \Rightarrow \mathrm{C}\neg\delta(f)_t$$

Notice that there is a significant difference between the encoding of rules about change from the encoding of rules about persistence: rules about change are causal, while persistence rules are not. Thus, while we will be able to explain a "clipping" in terms of a rule about change, we will not be able to explain the failure of a rule about change in terms of the persistence it fails to clip. Moreover, a persistence assumption is assumed not *applicable* when an event has caused the negation of the projected fluent. Thus, in particular, when an action causes a fluent f to be false at time t, the action is guaranteed not to affect the status of f prior to t. Notice also, that unlike the formulations of the Yale shooting problem discussed, this encoding scheme does not rely on making explicit the 'clippings' an action is expected to produce (e.g. $\mathtt{shoot}' \wedge \mathtt{alive}' \Rightarrow \neg\delta_2$). These 'clippings' will be inferred from the encoding above, according to the context in question.

Finally, cancellation rules stating that a given action is not applicable in a given circumstance are encoded in the form of causal rules. For instance, to state that the rule about change above is not applicable when some abnormal condition $abcond(x)$ holds, we would write the causal sentence $abcond(x)_t \Rightarrow \mathrm{C}\neg\delta_i(x)_t$.

We have illustrated above a causal encoding of a simple propositional version of the Yale shooting problem. A full temporal version would proceed along similar lines, rendering the same behavior. We consider now a slightly richer example due to Ginsberg and Smith [1988].

Example 5.10 Let us assume that there is a room with some ducts that maintain the room ventilated. If the ducts become blocked, the room becomes stuffy. Furthermore, an object sitting on a duct, blocks the duct. That is, we have three relevant causal relations:

$\mathtt{duct}(x) \wedge \exists y.\, \mathtt{on}(y,x)_t$	**causes**	$\mathtt{blocked}(x)_t$
$[\forall x.\mathtt{duct}(x) \Rightarrow \mathtt{blocked}(x)_t]$	**causes**	$\mathtt{stuffy}_{t+1}$
$\mathtt{move\_to}(x,y)_t$	**causes**	$\mathtt{on(x,y)}_{t+1}$

According to the guidelines sketched above, we encode these expressions in a background context K with sentences:[7]

$$\mathtt{duct}(x) \wedge \exists y.\, \mathtt{on}(y,x)_t \Rightarrow \mathrm{C}\mathtt{blocked}(x)_t$$
$$[\forall x.\mathtt{duct}(x) \Rightarrow \mathtt{blocked}(x)_t] \Rightarrow \mathrm{C}\mathtt{stuffy}_{t+1}$$
$$\mathtt{move\_to}(x,y)_t \wedge \delta_{\mathtt{move}}(x,y)_t \Rightarrow \mathrm{C}\mathtt{on}(x,y)_{t+1}$$
$$\mathtt{move\_to}(x,y)_t \wedge \mathrm{C}\neg\mathtt{on}(x,y)_{t+1} \Rightarrow \mathrm{C}\neg\delta_{\mathtt{move}}(x,y)_t$$

and a default schema $\mathtt{move\_to}(x,y)_t \rightarrow \delta_{\mathtt{move}}(x,y)_t$. The use of the predicate $\delta_{\mathtt{move}}$ in place of δ_i, for some integer i, is only for descriptive purposes.

The persistence of the fluents $\mathtt{blocked}(x,y)$, $\mathtt{on}(x,y)$, and $\mathtt{stuffy}$, and their negations, is expressed as stipulated above. To keep in mind that all these fluents are really *terms*,[8] we will use the notation $\overline{f}$ to denote the fluent which is the complement of f. Thus, for instance, $\overline{\mathtt{on(a,b)}}$ will stand for the 'negation' of $\mathtt{on(a,b)}$. Namely, if $\mathtt{on(a,b)}$ holds at time t, $\overline{\mathtt{on(a,b)}}$ will not, and vice versa. This is expressed by a constraint

$$f_t \Rightarrow \neg\overline{f}_t$$

which renders f and $\overline{f}$ incompatible. The complement of $\overline{f}$ is f itself.

[7] For simplicity, we treat some of the causal relations as strict. No significant change would arise, however, from treating them defeasibly.

[8] Recall that $\mathtt{blocked}(x,y)_t$ is an abbreviation of the *atom* $\mathrm{Holds}(\mathtt{blocked}(x,y),t)$.

Finally, for this example, we need to express that an object cannot be on two different places at the same time:

$$\mathtt{on}(x,y)_t \wedge \mathtt{on}(x,z)_t \Rightarrow y = z$$

Given this background context K, we consider a scenario $T = \langle K, E\rangle$, describing a room with two ducts $\mathtt{d_1}$ and $\mathtt{d_2}$. Furthermore, at time $\mathtt{t} = 0$ it is known that the room is not stuffy, that a block $\mathtt{a}$ is sitting on top of duct $\mathtt{d_1}$, and that a block $\mathtt{b}$ sitting on a place different than $\mathtt{d_2}$. Namely,

$$E = \{\mathtt{duct}(x) \Rightarrow x = \mathtt{d_1} \vee x = \mathtt{d_2}, \overline{\mathtt{stuffy}}_0, \mathtt{on}(\mathtt{a},\mathtt{d_1})_0, \overline{\mathtt{on}(\mathtt{b},\mathtt{d_2})}_0\}$$

Figure 5.5 depicts the situation and the backward and forward projections which are legitimized.

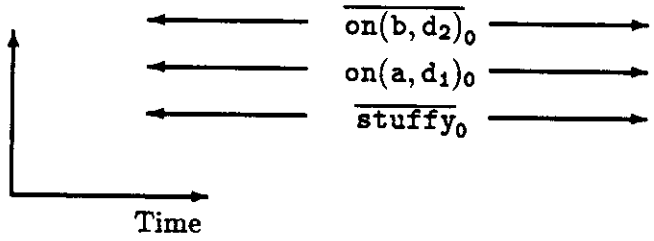


Figure 5.5: Initial scenario: $\overline{\mathtt{stuffy}}_0$, $\mathtt{on}(\mathtt{a},\mathtt{d_1})_0$, and $\overline{\mathtt{on}(\mathtt{b},\mathtt{d_2})}_0$

Let us now assume that at $\mathtt{t} = 0$ the block $\mathtt{d_2}$ is moved to duct $\mathtt{d_2}$. The action $\mathtt{move\_to}(\mathtt{b},\mathtt{d_2})$ determines a new context $T' = \langle K, E'\rangle$, with $E' = E + \{\mathtt{move\_to}(\mathtt{b},\mathtt{d_2})_0\}$ which produces two conflicts among assumptions (fig. 5.6).

These conflicts give rise to three minimal classes: the intended class C where the action is successful and, as a result, the two ducts get blocked and the room becomes stuffy; the class C', where the action is successful but somehow the block $\mathtt{a}$ has been removed from duct $\mathtt{d_1}$; and the class, C'', where the action is unsuccessful and the block $\mathtt{b}$ remains in a place different than $\mathtt{d_2}$. More precisely, the assumptions violated in each of these minimal classes are:[9]

$$\Delta[C] = \{\delta(\overline{\mathtt{stuffy}})_0, \delta(\overline{\mathtt{on}(\mathtt{b},\mathtt{d_2})})_0\}$$

[9]Recall that $\delta(f)_t$ stands for the assumption associated to the persistence of fluent f from t to $t+1$.

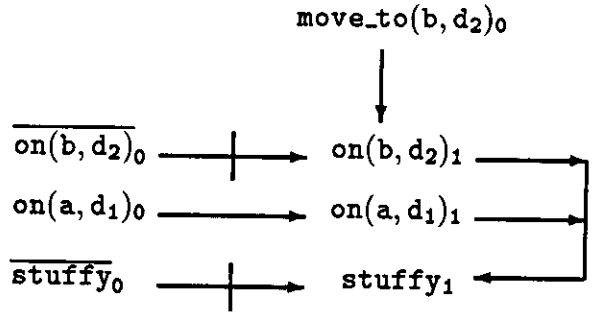


Figure 5.6: Scenario after moving block b to duct d_2

$$\Delta[\mathcal{C}'] = \{\delta(\mathtt{on}(\mathtt{a},\mathtt{d}_1))_0, \delta(\overline{\mathtt{on}(\mathtt{b},\mathtt{d}_2)})_0\}$$
$$\Delta[\mathcal{C}''] = \{\delta_{\mathtt{move}}(\mathtt{b},\mathtt{d}_2)_0\}$$

and each class explains the following violations:

$$\Delta^c[\mathcal{C}] = \{\delta(\overline{\mathtt{stuffy}})_0, \delta(\overline{\mathtt{on}(\mathtt{b},\mathtt{d}_2)})_0\}$$
$$\Delta^c[\mathcal{C}'] = \{\delta(\overline{\mathtt{on}(\mathtt{b},\mathtt{d}_2)})_0\}$$
$$\Delta^c[\mathcal{C}''] = \emptyset$$

As a result, $\mathcal{C}$ turns out to be the single causally preferred class of T, capturing the intuition that the block a stays on duct d_1 and that the room becomes stuffy. Note that such a behavior arises without the presence of explicit cancellation axioms.

5.3.3 Logic Programming

While the adequacy of the framework presented for reasoning about change and inheritance hierarchies rests mainly on empirical grounds —how natural it is to express knowledge about these domains and how closely the resulting behavior resembles the behavior intended by the user— a growing body of work on the semantics of general logic programs will permit us to assess the expressivity and semantics of causal theories on formal grounds.[10]

[10]For a review of the different semantic account of logic programs, see [Shepherson, 1987] and [Przymusinska and Przymusinski, 1989]. For an introduction to logic programming, see Lloyd [1984].

General logic programs are collections of implicitly universally quantified rules of the form: $A \leftarrow L_1, L_2, \ldots, L_n$, where A is an atom called the head of the rule, and each L_i, $i = 1, \ldots, n$, $n \geq 0$, is a positive or negative literal in the rules' body. When all the literals L_i are positive, the rule is said to be positive. Logic programs composed only of positive rules are called *positive* logic programs.

The interest in logic programming was sparkled by the development of Prolog, a general purpose programming language with close connections to the Horn-subset of classical first order logic [Rousell, 1975]. The Prolog experience showed that it was possible to combine the declarative reading of logic with the procedural interpretation of more conventional programming languages [Kowalski, 1979]. A positive rule such as $A \leftarrow L_1, L_2, \ldots, L_n$ can thus be understood both as stating that A *is* true when all the literals L_i, $i = 1, \ldots, n$ are true, and that the goal A is derivable when each subgoal L_i is derivable. If all the rules are positive and the connective '$\leftarrow$' is regarded as material implication, it can be shown that the set of ground atoms true in all models of a *positive* program P corresponds precisely to the set of atoms true in the single *minimal* Herbrand model of P. Such minimal Herbrand model is identical in turn, to the set of ground atoms which can be derived from P by a careful form of back-chaining called SLD-resolution [Lloyd, 1984].

When some of the literals L_i are negative, however, things are not so simple and the declarative reading of logic programs is usually dropped. Such programs are commonly understood in procedural terms, with the proviso that negative literals $\neg A_i$ are assumed derivable when every derivation for A_i fails. Such an interpretation of negation has turned out to be a particularly useful programming tool, and follows a tradition that goes back to Planner-like languages [Hewitt, 1972]. Coined *negation as failure*, it endows logic programs with a behavior that is non-monotonic. In a program containing a single rule $\mathtt{p} \leftarrow \neg\mathtt{q}$, for instance, negation as failure yields a derivation for the atom p, which no longer holds when a rule such as $\mathtt{q} \leftarrow$ is added.

The first logical interpretation of the negation as failure rule is due to Clark [1978]. While under the standard interpretation of positive logic programs, the collection of rules of the form $A \leftarrow L_1, \ldots, L_n$ for a given atom A, are regarded to provide *sufficient* conditions for the atom A to hold, under Clark's interpretation, these conditions are assumed to be *necessary* as well. Thus, for instance, the semantics of a program P comprised of the rules $\mathtt{p} \leftarrow \neg\mathtt{q}$ and $\mathtt{q} \leftarrow \mathtt{r}$ is identified with the semantics of the *completion* of P, written comp[P], given by the collection of sentences $\mathtt{p} \Leftrightarrow \neg\mathtt{q}$ and $\mathtt{q} \Leftrightarrow \mathtt{r}$ and $\mathtt{r} \Leftrightarrow$ **false**. From comp[P], the truth of p and the falsehood of q and r do indeed follow, in agreement with negation as failure.

Assuming that the user means comp[P], when s/he writes the program P, Clark and others have shown the negation as failure rule to be sound and complete for a large class of programs. Nonetheless, there are still important limitations in Clark's approach, even in the domain of positive programs. For instance, the completion of a program $\mathtt{p} \leftarrow \mathtt{p}$ fails to sanction the proposition $\neg\mathtt{p}$, even though it holds in its unique minimal Herbrand model.[11] Furthermore, the addition of such an apparently innocent rule to a program $\mathtt{q} \leftarrow \neg\mathtt{p}$, prevents Clark's interpretation from sanctioning the otherwise expected conclusion q.

In recent years a number of new semantic accounts of general logic programs have been proposed (see the reviews in [Shepherson, 1987, Przymusinska and Przymusinski, 1989]) Not only do these new accounts overcome the shortcomings of Clark's approach, but they also suggest interesting connections with other non-monotonic formalisms proposed in AI (e.g., [Przymusinski, 1988]). We will not review these approaches here; rather we will compare them with the result of interpreting logic programs as particular types of causal theories.

A Causal Semantics for Logic Programs

Let $\mathcal{L}$ be a first order language. The *Herbrand universe* of $\mathcal{L}$, $U_{\mathcal{L}}$, is the set of all ground *terms* in $\mathcal{L}$, while the *Herbrand base* of $\mathcal{L}$, $B_{\mathcal{L}}$, is the set of all ground *atoms* in $\mathcal{L}$. A *Herbrand interpretation* over $\mathcal{L}$ is an interpretation whose domain is the Herbrand universe $U_{\mathcal{L}}$, where each constant symbol is assigned to itself, and each function symbol f^n is assigned to the mapping $(t_1, t_2, \ldots, t_n) \mapsto f^n(t_1, t_2, \ldots, t_n)$ from $(U_{\mathcal{L}})^n$ to $U_{\mathcal{L}}$. A Herbrand interpretation can be represented by the set of atoms in the Herbrand base that it satisfies. For specifying the models of a program P, the connective '$\leftarrow$' is interpreted as material implication, and rules are assumed to be universally quantified. Rules of the form $A \leftarrow$ are interpreted as $A \leftarrow \mathbf{true}$, where **true** is an special atom satisfied in every interpretation. A *Herbrand model* of P is a Herbrand interpretation that satisfies all the rules in P.

As it is standard, we consider only the Herbrand models of programs. Moreover, since for answering existential queries, a program involving variables can be shown to be equivalent to a program without variables,[12] we will be dealing mainly with variable-free logic programs. More precisely, we will analyze the semantics of general logic programs in terms of three different mappings $C_i[\cdot]$, $i = 1, 2, 3$, each

[11] Note, however, that Clark's interpretation remains close to the semantics of the negation as failure rule: a proof for p will normally 'loop' in such program without ever returning failure.

[12] This is a consequence of Herbrand's theorem (see for instance, [Chang and Lee, 1973].

converting a program P over a language $\mathcal{L}$ into a causal theory $C_i[P]$ over the language $\mathcal{CL}$. Each such mapping associates a different "meaning" to P. For the purposes of logic programming, $C_2[P]$ will turn out to be most relevant, as it represents an extension of Przymusinski's [1987] perfect model semantics. The alternative mappings $C_1[\cdot]$ and $C_3[\cdot]$ will be used mainly to illustrate the relation between the interpretation of general logic programs and the semantics of causal theories.

Each causal theory $C_i[P]$, $i = 1, 2, 3$ takes the form $C_i[P] = \langle K_i[P], E \rangle$, with a background $K_i[P] = \langle L_i[P], D \rangle$, where the set of defaults D and the evidence set E are invariably empty. Thus, the minimal classes and the conditional classes of $C_i[P]$ coincide. The only sentences in $C_i[P]$ then, are just those in the background context which correspond to causal rules obtained by an appropriate local transformation of the rules in the logic program P.

The assumptions δ in the causal theories $C_i[P]$ will correspond to the set of *negative* literals in the underlying (non-causal) language $\mathcal{L}$. As a matter of convenience, for a given theory $C_i[P]$, $\mathcal{C}_A$ will denote the class of Herbrand models of $C_i[P]$ whose plain atoms (i.e. exceptions) are among those of A. We will say that A stands for the *atomic* gap of the class $\mathcal{C}_A$, to distinguish it from the set Δ of assumptions violated by some model in $\mathcal{C}_A$, which remains as the gap of the class. We will further convene that an assumption $\delta = \neg p$ in Δ belongs to the explained gap of $\mathcal{C}_A$, when the causal literal $\mathrm{C}p$ (rather than $\mathrm{C}\neg\neg p$) holds in the class. This will permit us to dispense with the constraints [C2] and [C3] on the operator C, and only retain [C1], which requires models of $\mathrm{C}p$ to be models of p as well.

We consider first the mapping $C_1[\cdot]$ which associates programs P with causal theories of the form $C_1[P]$, by converting each rule

$$\gamma \leftarrow \alpha_1, \ldots, \alpha_n, \neg\beta_1, \ldots, \neg\beta_m \quad ,$$

in P, where $n \geq 0$ and $m \geq 0$, and α's, β's and γ are atoms, into a *causal* rule of the form

$$\mathrm{C}\alpha_1 \wedge \ldots \wedge \mathrm{C}\alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow \mathrm{C}\gamma \quad .$$

The example below illustrates such a translation, together with the semantics of the original program P and the resulting causal theory $C_1[P]$.

Example 5.11 Consider a program P given by the following rules:

$$\mathsf{c} \leftarrow \mathsf{a}, \neg\mathsf{b}$$
$$\mathsf{d} \leftarrow \neg\mathsf{c}$$
$$\mathsf{a} \leftarrow$$

This program has two minimal models: $M_1 = \{\mathsf{a}, \mathsf{c}\}$ and $M_2 = \{\mathsf{a}, \mathsf{b}, \mathsf{d}\}$. The model M_1, however, is the single *canonical* model of P [Apt *et al.*, 1987].

The mapping $C_1[\cdot]$ maps the program P into the causal theory $C_1[P]$, given by the following causal rules:

$$\mathsf{C}\mathsf{a} \wedge \neg\mathsf{b} \Rightarrow \mathsf{C}\mathsf{c}$$
$$\neg\mathsf{c} \Rightarrow \mathsf{C}\mathsf{d}$$
$$\mathbf{true} \Rightarrow \mathsf{C}\mathsf{a}$$

Such a causal theory gives rise to two minimal classes $\mathcal{C}_{M_1}$ and $\mathcal{C}_{M_2}$, with atomic gaps M_1 and M_2 as above. Furthermore, in the former class, the atoms a and c are both explained, as $C_1[P], \neg\mathsf{b} \vdash \mathsf{C}\mathsf{a} \wedge \mathsf{C}\mathsf{c}$ holds, and $\neg\mathsf{b}$ is a legitimate assumption in $\mathcal{C}_{M_1}$. On the other hand, only the atom a is explained in $\mathcal{C}_{M_2}$. Thus, the class $\mathcal{C}_{M_1}$, which is a perfectly coherent class, is the single causally preferred class of $C_1[P]$. As a result, the canonical model M_1 of P and the causally preferred class $\mathcal{C}_{M_1}$ of $C_1[P]$ sanction the same non-causal literals.

This example suggests a possible correspondence between the 'intended' behavior of a logic program P and the interpretation of the causal theory $C_1[P]$. Such a correspondence can be specifically tested in the class of *stratified* programs, whose intended behavior has been formalized in various ways. Stratified programs are general logic programs in which the use of negation adheres to certain constraints. These constraints permit to characterize a stratified programs P in terms of a single minimal Herbrand model, called either the *canonical* of P [Apt *et al.*, 1987]. the *perfect* model of P [Przymusinski, 1987], the *stable* model of P [Gelfond and Lifschitz, 1988], or the *felicitous* model of P. [Fine, 1989]. In the case of propositional programs, a program P is stratified when its rules can be organized in layers, in such a way that rules in which a literal $\neg p$ occurs in their bodies appear in higher layers than rules in which the atom p occurs in their head.[13] For such

[13]Similarly, a dependency graph of the program P can be constructed in which a link between atoms p and q occurs when there is a rule with head q and a body which includes p. The link is positive if the atom p occurs positively in the body, and negative if the atom p occurs negatively in the body [Apt *et al.*, 1987]. P is then stratified if and only if there are no cycles containing negative links in the dependency graph of P.

programs, the following correspondence between the canonical model of P and the single causally preferred class of the theory $C_1[P]$ can be established:

Theorem 5.1 *Let P be a stratified program. Then M is the canonical model of P if and only if $\mathcal{C}_M$ is the single causally preferred class of $C_1[P]$.*

As mentioned above, this implies that the *non-causal literals* sanctioned by M and $\mathcal{C}_M$ are the same. Models in $\mathcal{C}_M$, however, will normally satisfy additional *causal* literals. Indeed, $\mathcal{C}_M$ is a perfectly coherent class, and thus if a plain atom p belongs to M, the causal literal Cp will belong to every model in $\mathcal{C}_M$.

The perfectly coherent classes of the causal theory $C_1[P]$ are closely related to the *stable* models of the program P, *even when P is not stratified*. Namely, when M is a stable model of an *arbitrary* program P, the class $\mathcal{C}_M$ is a perfectly coherent class of P, and vice versa:[14]

Lemma 5.1 *M is a* stable *model of an arbitrary program P if and only if $\mathcal{C}_M$ is a perfectly coherent class of the causal theory $C_1[P]$.*

However, even in light of this correspondence, the semantics of causal theories $C_1[P]$ and the stable semantics of logic programs P diverge outside the family of non-stratified programs. The reasons for such a departure are several.

First, even when a program P may have no stable models, the associated casual theory $C_1[P]$ may have a set of non-perfectly coherent causally preferred classes. The simplest such an example is the program P_0 given by the single rule $\mathtt{p} \leftarrow \neg\mathtt{p}$. Such a program does not admit stable models, while the causal theory $\neg\mathtt{p} \Rightarrow \mathrm{C}\mathtt{p}$ has a single causally preferred class $\mathcal{C}_M$, with $M = \{\mathtt{p}\}$. As a result, a query such as p? will remain undefined in a stable model semantics, but will be answered positively under the causal interpretation described.[15]

While the program P_0 has no stable models, other programs may have multiple stable models. For instance the program P_1 given by the rules $\{\mathtt{a} \leftarrow \neg\mathtt{b},\ \mathtt{b} \leftarrow$

[14]Readers not familiar with Gelfond and Lifschitz's [1988] stable semantics of logic programs, may want to adopt this correspondence as the 'definition' of the stable semantics for the sake of the discussion below. Gelfond and Lifschitz's [1988] *stable* models are also equivalent to Fine's [Fine, 1989] *felicitous* models.

[15]There is no consensus in the logic programming community about the 'right' answer to this query. Approaches based on 3-valued models such as Van Gelder *et al.* [1988] and Przymusinski [1989] will also leave the value of p undefined.

$\neg$a, p $\leftarrow$ $\neg$a} has two stable models: $M_1 = \{\mathtt{a}\}$ and $M_2 = \{\mathtt{b}, \mathtt{p}\}$. Likewise, the causal encoding $C_1[P]$ of P has two causally preferred classes $\mathcal{C}_{M_1}$ and $\mathcal{C}_{M_2}$. Nonetheless, differences arise again, when the programs P_0 and P_1 are combined into a single program P_2. As noted by Van Gelder *et al.* [1988], the addition of the rules in P_1 to P_0 has a stabilizing effect; and while P_0 does not possess stable models, the program P_2 does. Indeed, $M = \{\mathtt{b}, \mathtt{p}\}$ turns out to be the *unique* stable model of P. On the other hand, the causal theory $C_1[P_2]$ retains *two* causally preferred classes: the class $\mathcal{C}_M$, with M as above, and the class $\mathcal{C}_{M'}$, with $M' = \{\mathtt{a}, \mathtt{p}\}$.

The differences in these examples, nonetheless, are justifiable, and arguably, even advantageous. In other cases, however, such a departure is less convincing. The problem in ascribing a semantics to a program P in terms of the "atomic gaps" of the causally preferred classes of the transformed program $C_1[P]$, is that not always these atomic gaps are models of P.

As an example, consider the program $P_3 = \{\mathtt{p} \leftarrow \neg\mathtt{p},\ \mathtt{q} \leftarrow \mathtt{p}\}$. The causal theory $C_1[P_3]$ associated with P_3 is given by the causal rules $\{\neg\mathtt{p} \Rightarrow \mathrm{C}\mathtt{p},\ \mathrm{C}\mathtt{p} \Rightarrow \mathrm{C}\mathtt{q}\}$. Since there are interpretations that simultaneously satisfy the literals p, $\neg$Cp and $\neg$q, the single causally preferred class of $C_1[P_3]$ is $\mathcal{C}_M$, with $M = \{\mathtt{p}\}$. As a result, the literals p and $\neg$q are sanctioned by this causal semantics, even though there is no model of P in which p holds and q does not.

Thus, while the mapping $C_1[\cdot]$ provides a satisfactory interpretation of stratified logic programs, it is inadequate for a large class of non-stratified programs. In particular, as the program P_3 demonstrates, such an interpretation fails to sanction certain *logical* consequences of the program under consideration. We will show now that an alternative mapping $C_2[\cdot]$ of logic programs into causal theories, removes these difficulties while succeeding in retaining the satisfactory features of $C_1[\cdot]$.

The mapping $C_1[\cdot]$ considered above, translates each rule

$$\gamma \leftarrow \alpha_1, \ldots, \alpha_n, \neg\beta_1, \ldots, \neg\beta_m$$

in a program P, into a *causal* rule of the form

$$\mathrm{C}\alpha_1 \wedge \ldots \wedge \mathrm{C}\alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow \mathrm{C}\gamma$$

in the theory $C_1[P]$. The new mapping $C_2[\cdot]$ is defined in a similar manner, except that the causal rules are supplemented by the direct logical encoding of the rules in P:

$$\alpha_1 \wedge \ldots \wedge \alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow \gamma$$

The theory $C_2[P]$ is thus the composition of $C_1[P]$ and P. The effect of extending the causal rules of $C_1[P]$ with the rules of the program P itself, is to prune those models of $C_1[P]$ which are not models of the original program; the causal theory $C_1[P]$ thus determines the "preferences" on models, while P remains as an 'integrity constraint.' As a result, the "atomic gaps" of the causally preferred classes C_M of $C_2[P]$ are now guaranteed to be models of the target program P, suggesting the following definition:

Definition 5.3 *M is an* induced causal model *of a program P iff C_M is a causally preferred class of $C_2[P]$.*

The previous result about stratified programs can therefore be cast as follows:

Theorem 5.2 *For a stratified program P, there is a single induced causal model which is identical to the canonical model of P.*

We will refer to the semantics of P determined by its induced causal models, as the *causal semantics* of P. Note that while the mappings $C_1[\cdot]$ and $C_2[\cdot]$ induce an identical interpretation of *stratified* logic programs, $C_2[\cdot]$ is better suited than $C_1[\cdot]$ for non-stratified programs. The semantics induced by the mapping $C_2[\cdot]$ can indeed be regarded as an extension Przymusinski's perfect model semantics in the domain of non-stratified programs. The appeal to the —em dynamic notion of "explanation" in the interpretation of causal theories, provides in fact an ordering on atoms similar to that obtained by the extension of the perfect model semantics proposed in [Przymusinska and Przymusinski, 1988]. We illustrate this in the following example.

Example 5.12 Consider the following rules

(1) $\mathtt{b} \leftarrow$

(2) $\mathtt{d} \leftarrow$

(3) $\mathtt{a} \leftarrow \mathtt{b}, \neg\mathtt{c}$

(4) $\mathtt{c} \leftarrow \mathtt{d}, \neg\mathtt{a}$

and two non-stratified programs $P_{2,3,4}$ and $P_{1,3,4}$, comprising rules 2,3,4 and 1,3,4 respectively. Both $P_{2,3,4}$ and $P_{1,3,4}$ contain a rule which appears to be 'irrelevant'

in the context of the other rules in the program: rule (3) appears irrelevant in $P_{2,3,4}$ as no information is available about its positive antecedent b, while rule (4) appears irrelevant in $P_{1,3,4}$ as no information is available about d. If the 'irrelevant' rules are removed from each program, two *stratified* programs $P_{2,4}$ and $P_{1,3}$ result. In particular, the first program legitimizes the atom a, while the second program legitimizes the atom c. If the rules are not removed from the original programs, however, neither of these atoms is legitimized by the perfect model semantics. Indeed, for sanctioning such a behavior it is necessary to determine the priorities relating the atoms a and c dynamically; i.e. by looking at the program as a whole, rather than at its individual rules. This, however, is beyond the power of perfect model semantics, which relies on a priority ordering on atoms established by considering each rule in isolation. The causal semantics described has no such a limitation. The assumption $\neg$a is "preferred" to the assumption $\neg$c in the context of the causal theory $C_2[P_{2,3,4}]$, while the assumption $\neg$c is preferred to the assumption $\neg$a in the context of $C_2[P_{1,3,4}]$.[16] By determining "preferences" in terms of *global* explanations, the causal interpretation of logic programs thus remains unaffected by the presence of "irrelevant" rules.

Logic Programs and Causal Networks

We have considered so far the semantics of causal theories that result from mapping the rules

$$\gamma \leftarrow \alpha_1, \ldots, \alpha_n, \neg\beta_1, \ldots, \beta_m$$

of a logic program P, into causal rules of the form

$$\mathrm{C}\alpha_1 \wedge \ldots \wedge \mathrm{C}\alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow \mathrm{C}\gamma \quad .$$

Causal theories $C_2[P]$ contain, in addition, the logical encoding of the program P itself. Both mappings yield a satisfactory interpretation of stratified logic programs, and $C_2[\,\cdot\,]$, in particular, yields an appealing interpretation of non-stratified programs as well. We will now investigate the semantics associated with a third mapping $C_3[\,\cdot\,]$ of logic programs into causal theories. The interest on this mapping does not lie on what it has to say about general logic programs or causal theories,

[16] For an understanding of causal entailment in terms of priorities, see section 6.2 below.

but, rather, for what it suggests about the relation between *logic programs* and *causal networks.*

For a logic program P, $C_3[P]$ represents the collection of rules which result from mapping each rule

$$\gamma \leftarrow \alpha_1, \ldots, \alpha_n, \neg\beta_1, \ldots, \neg\beta_m$$

in P, into a causal rule of the form:

$$\alpha_1 \wedge \ldots \wedge \alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow \mathrm{C}\gamma$$

The difference with the previous translations is that the positive antecedents of the resulting causal rules do not need to be "causally" established. This renders the models of the causal theory $C_3[P]$ as models of the causal theory $C_2[P]$ though not the other way around. The semantics of causal theories of the form $C_3[P]$ differs from that of $C_2[P]$ and $C_1[P]$ even within the family of stratified programs P. We illustrate this departure in the example below.

Example 5.13 Let P be the program given by the following rules:

$$\mathtt{q} \leftarrow \neg\mathtt{p}$$
$$\mathtt{p} \leftarrow \mathtt{r}$$
$$\mathtt{r} \leftarrow \mathtt{p}$$

P is stratified, and therefore, possesses a single canonical model $M = \{\mathtt{q}\}$. In light of the results above, thus, $\mathcal{C}_M$ is the single causally preferred class of the theories $C_1[P]$ and $C_2[P]$.

The mapping $C_3[\,\cdot\,]$, on the other hand, renders the following causal theory:

$$\neg\mathtt{p} \Rightarrow \mathrm{C}\mathtt{q}$$
$$\mathtt{r} \Rightarrow \mathrm{C}\mathtt{p}$$
$$\mathtt{p} \Rightarrow \mathrm{C}\mathtt{r}$$

Such a theory accepts two minimal classes, $\mathcal{C}_M$ and $\mathcal{C}_{M'}$, whose atomic gaps $M = \{\mathtt{q}\}$ and $M' = \{\mathtt{p}, \mathtt{r}\}$, are in correspondence with the two minimal models of the original program P. Furthermore, both classes are perfectly coherent, and thus, both are causally preferred: the assumption $\neg\mathtt{q}$ explains p in the class $\mathcal{C}_M$, while the atoms p and q explains each other in the class $\mathcal{C}_{M'}$. As a result, the theories $C_1[P]$ and $C_2[P]$ support the truth of q, while the theory $C_3[P]$ does not.

It is clear in this example that the 'anomalous' behavior of the theory $C_3[P]$ is a consequence of the circularity relating the atoms p and r. 'Circular' explanations are precluded in the theories $C_1[P]$ and $C_2[P]$, in which the assumption $\neg$q supports the truth of p and r but not of Cp or Cr. What is interesting, however, is that once these circularities are removed, the 'anomalous' behavior is guaranteed to disappear.

Let us say that a program P is *acyclic* when its dependency graph does not contain cycles. Acyclic programs are thus stratified. Moreover, acyclic programs, not only preclude 'recursion trough negation,' but *every* type of recursion. For acyclic programs, the following result applies.

Theorem 5.3 *Let P be an acyclic program. Then the class C_M, where M is the canonical model of P, is the unique causally preferred class of the theories $C_1[P]$, $C_2[P]$ and $C_3[P]$.*

In other words, once recursion is removed the three causal mappings examined result into an identical behavior, in correspondence with the received semantic accounts of logic programs. While the requirement of acyclicity is unacceptably strong in the domain of programming, it is common among network representational languages, such as inheritance hierarchies [Touretzky, 1986] and Bayesian networks [Pearl, 1988b]. Indeed, causal theories of the form $C_3[P]$, for acyclic programs P, possibly augmented by integrity constraints[17] and non-assumption negative literals, provide a sufficiently expressive language for reasoning in *causal networks*. Pearl's [1988b, section 4.3.2] *noisy-OR* probabilistic networks, for example, consist essentially of rules of the form

$$\alpha_1 \wedge \ldots \wedge \alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow \mathrm{C}\gamma$$

where α_i's are *preconditions* and β_i's are *censors*, which are disjunctively combined into the common 'node' γ. We will investigate the use of such a representational language for diagnostic reasoning in section 5.3.4. Below, we illustrate some of the differences between the semantics of theories $C_3[P]$ and $C_2[P]$ for a program P containing negative clauses.

[17] By integrity constraints we mean arbitrary propositional sentences which do not involve the causal operator C.

Example 5.14 Consider the acyclic logic program P given by the rules:

$$\mathtt{h} \leftarrow \mathtt{s}, \neg\mathtt{v}$$
$$\mathtt{m} \leftarrow \mathtt{v}, \neg\mathtt{e}$$
$$\mathtt{s} \leftarrow$$

Let us say that the first rule asserts that "whenever Tom needs to go to the supermarket (s) he goes to Hughes (h), unless he goes to Vons (v)," and the second rule that "whenever he goes to Vons, he buys mangos[18] (m), unless he has enough already (e)."

In the context of P, all the accounts considered point to a single preferred model: $M = \{\mathtt{s}, \mathtt{h}\}$. Now, let us consider that we learn the additional information that Tom is not only going to the supermarket (s), but that, this time, he is not going to Hughes:

$$\leftarrow \mathtt{h}$$

Intuition dictates that, in the new context, there should be a single preferred model $M' = \{\mathtt{s}, \mathtt{v}, \mathtt{m}\}$. That is, since Tom goes to Hughes except when he goes to Vons, he must be going to Vons (v) now that he is not going Hughes. Furthermore, since when he goes to Vons he buys mangos except when he has enough, he must be buying mangos (m) too.

Nonetheless, neither the stable model semantics nor Clark's completion semantics apply to this new 'program', as there is no stable model of P which satisfies $\neg$h, nor a completion of P consistent with $\neg$h. Furthermore, both causal theories $C_1[P]$ and $C_2[P]$ admit two preferred classes: $\mathcal{C}_{M'}$ and $\mathcal{C}_{M''}$, with $M'' = \{\mathtt{s}, \mathtt{v}, \mathtt{e}\}$, leaving the status of the atoms m and e undecided.

In contrast, the intended behavior is captured by the causal theory $C_3[P]$:

$$\mathtt{s} \wedge \neg\mathtt{v} \Rightarrow \mathrm{C}\mathtt{h}$$
$$\mathtt{v} \wedge \neg\mathtt{e} \Rightarrow \mathrm{C}\mathtt{m}$$
$$\mathbf{true} \Rightarrow \mathrm{C}\mathtt{s}$$
$$\neg\mathtt{h}$$

The classes $\mathcal{C}_{M'}$ and $\mathcal{C}_{M''}$ both remain as the minimal classes, but only the former is now causally preferred. The reason for such a preference is that the explanation of m no longer requires an explanation for v.

[18] A tropical fruit.

5.3.4 Abductive Reasoning

A leading motivation behind the work in non-monotonic reasoning has been the goal of providing a formal account of some of the pervasive patterns of commonsense inference. So far, our focus has been on default inference; a form of reasoning akin to deductive inference, where assumptions are adopted in the absence of contrary evidence. Nonetheless, there are other forms of non-monotonic inference, qualitatively different from default reasoning, which also appear to play an important role in commonsense reasoning. One such form, is what has been referred to either as "inference to the best explanation,"" abductive reasoning," or "conjectural reasoning" [Peirce, 1955, Harman, 1986, Charniak and McDermott, 1985]. This is a form of inference which attempts to make sense of the evidence when it does not cohere with a given set of beliefs. The characterization of these patterns of inference involves both the determination of the sources of incoherence and the identification of hypotheses capable of explaining such incoherence away. In this subsection, we will to show that the framework we have so far developed, lends naturally to a characterization of that sort.

The central idea is too associate a *coherence* measure to *contexts* as opposed to *classes* of models. Intuitively, the coherence —or, for that matter, the incoherence— of a context T will depend on whether the 'exceptions' declared in that context have an explanation. A natural choice is to associate to T an *incoherence set* in which the *unexplained gaps* of its preferred classes are grouped. In particular, such an incoherent set will be empty if the preferred classes of T are perfectly coherent; and non-empty, otherwise.

More precisely, if $\mathcal{C}_i$, $i = 1, \ldots, n$ are the preferred classes of T,[19] we define the *incoherence set* $I[T]$ of T, to be the collection of sets $\Delta^u[\mathcal{C}_i] = \Delta[\mathcal{C}_i] - \Delta^c[\mathcal{C}_i]$, $i = 1, \ldots, n$, where $\Delta[\mathcal{C}_i]$ and $\Delta^c[\mathcal{C}_i]$ stand for the gap and the explained gap of $\mathcal{C}_i$ respectively.

For instance, a context in which all that is known is that Tim is a non-flying bird will have a non-empty incoherence set, reflecting the lack of explanation for Tim's unexpected feature. On the other hand, if it is learned that Tim is sick, such incoherence would become explained, leaving the new context with an empty incoherence set.

We will say that a context T is *as coherent as* as a context T' if for every set S

[19]Unless otherwise specified, the preferred classes refer to the causally preferred conditional classes.

in $I[T]$ there is a set S' in $I[T']$ such that $S \subseteq S'$. If the context T is as coherent as T', but T' is not as coherent as T, we will further say that T is *more coherent* than T'.

In the example above, for instance, the context in which Tim is known to be a *sick* non-flying bird, is *more coherent* than the context in which all that is known is that Tweety is a non-flying bird.

In principle, it would be natural to say that a proposition p qualifies as a conjecture in a context T, if the context $T + \{p\}$ is more coherent than the context T;[20] namely, if the adoption of p in T renders previous unexplained exceptions explained without introducing new ones. The problem with this approach, however, is that it gives rise to too many conjectures. For instance, if both p and q qualify as conjectures in a context T, their disjunction $p \vee q$ will usually qualify as well. So, together with the conjectures 'Tim is sick' and 'Tim is a penguin', we would obtain the unnatural conjecture: "Tim is sick or Tim is a penguin." This proliferation of conjectures could in principle be avoided by imposing syntactic restrictions on the form of the admissible conjectures, though such a criterion will likely give rise to different conjectures for different conceptualizations of the domain of interest.[21]

Our approach here will be slightly different. Like Poole [1987], we will assume that conjectures are selected from a predetermined set, which we will refer as $\mathbf{\Xi}$. The difference with Poole, however, is that conjectures are going to be invoked to explain incoherence rather than observations. The latter task will turn out to be a special case of the former, when (certain) observations are declared to be abnormal.

A context T and a set of conjectures $\Xi \in \mathbf{\Xi}$ logically consistent with T, will determine what we call a *belief state* $\langle T, \Xi \rangle$. A belief state $\mathcal{B} = \langle T, \Xi \rangle$ is thus a context $T + \Xi$ in which part of the evidential component stands for hypothetical beliefs rather than solid evidence. We say in that case that the belief state $\mathcal{B}$ is *rooted* in T.

For two belief states $\mathcal{B} = \langle T, \Xi \rangle$ and $\mathcal{B}' = \langle T, \Xi' \rangle$, $\mathcal{B}$ is *less committed* than $\mathcal{B}'$, if $\Xi \subset \Xi'$, and $\mathcal{B}$ is a *maximally-coherent* belief state, if there is no other belief state $\mathcal{B}'$ rooted in T which is more coherent than $\mathcal{B}$.

Finally, a maximally-coherent belief state $\mathcal{B} = \langle T, \Xi \rangle$ is *admissible* when there

[20] For a context $T = \langle K, E \rangle$, we use $T + \{p\}$ to denote the context $T' = \langle K, E \cup \{p\} \rangle$.

[21] This is not necessarily bad though. The choice of predicates in the conceptualization of a body of knowledge is likely to contain information about the structure of the domain.

is no maximally-coherent belief state $\mathcal{B}' = \langle T, \Xi' \rangle$ which is less committed than $\mathcal{B}$. Intuitively, admissible belief states $\langle T, \Xi \rangle$ are supposed to represent maximally coherent belief states that a rational agent with the information in T may choose to adopt. We also say in that case that Ξ is an *admissible hypothesis set* in T, and that the conjectures in Ξ are *admissible hypotheses* in T.

We will illustrate these definitions with examples from Console *et al.* [1989], and Pearl [1988a]. As we will see, the fact that what needs to be explained and what constitutes an explanation are part of the language of causal theories, makes the present framework particularly suitable for applications involving reasoning from evidence to hypotheses.

Example 5.15 Let us consider the causal network depicted in fig. 5.7 describing a fragment of the knowledge relevant to the diagnosis of a malfunctioning car. We will encode such a network by mapping each causal link $\alpha \rightarrow \beta$ into a causal rule $\alpha \Rightarrow C\beta$, and by regarding each atom in the net as an 'exception' that needs to be explained. Furthermore, we assume a pool Ξ of conjectures which includes only the top propositions `pistons_rings_used`, `oil_cup_holed`, `old_spark_plugs`, which we regard as self-explanatory; namely for each such conjecture ξ we assume $\xi \Rightarrow C\xi$.

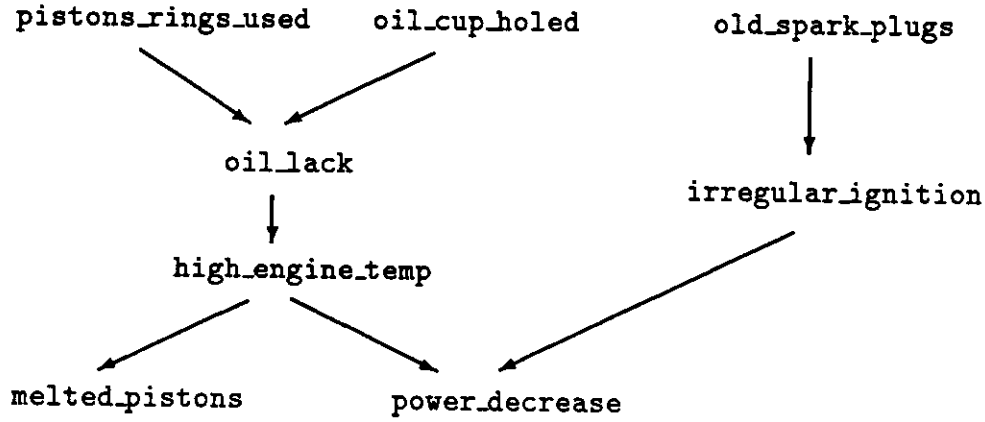


Figure 5.7: A causal network

Let us assume now that `power_decrease` is observed, and let T refer to the resulting context. Clearly there is a single preferred class of T in which the exception `power_decrease` holds but is not explained; namely, there is no set of *assumptions*

Δ consistent with T which supports C`power_decrease`. On the other hand, any belief state $\mathcal{B}_i = \langle T, \Xi_i \rangle$ with a non-empty set of *conjectures* Ξ_i, will explain such incoherence away, and thus every such state is *maximally-coherent.* However, only those states containing a *single* hypothesis from Ξ will qualify as *admissible* belief states. Thus, there are three admissible belief states, involving three singleton hypothesis sets {`pistons_rings_used`}, {`oil_cup_holed`}, and {`old_spark_plugs`} respectively.

If ¬`high_engine_temp` is also observed, however, only one admissible hypothesis would remain: {`old_spark_plugs`}.[22] Moreover, if we also add ¬ `irregular_ignition` no admissible conjecture would be left, thus giving rise to a single (incoherent) belief state involving no conjecture at all. Note that this behavior is different from the behavior that results from approaches which equate abduction with deduction in a *completed model* (e.g. [Kautz, 1987] and [Console *et al.*, 1989]). Indeed, the completed model when `power_decrease`, ¬`high_engine_temp` and ¬`irregular_ignition` have all been observed is inconsistent.

The framework for conjectural reasoning suggested also permits to accommodate what Pearl has called *evidential defaults.* In [Pearl, 1988a] Pearl suggested a distinction between defaults which evoke expectations from those which evoke explanations. He called the former defaults, much as we do it here, *causal defaults*, and the latter, *evidential defaults*, in analogy to the distinction between causal and evidential support in the context causal probabilistic networks [Pearl, 1988b, chap.4]. Thus, while a default such as "rain → grass-is-wet" represents a causal default, the converse default "grass-is-wet → rain" represents an evidential default. Pearl further argued that while causal defaults may trigger other causal defaults, a counterintuitive behavior is likely to result when causal defaults trigger evidential defaults. This explains why the reasoning chain "my-shoes-are-wet → the-grass-is-wet → it-rained" is sound, while a reasoning chain "the-sprinkler-was-on → the-grass-is-wet → it-rained" is not. We use Pearl's example below to show how evidential defaults can be expressed in the framework proposed.

Example 5.16 Consider the causal network depicted in fig. 5.8. We encode this network again by translating each link $\alpha \rightarrow \beta$ into a causal rule $\alpha \Rightarrow C\beta$. The 'modified' link `sprinkler_on` → `grass_wet`, on the other hand, is mapped into the rule `sprinkler_on` ∧ ¬`bad_pipes` ⇒ C`grass_wet`. All tokens in the network are

[22] Note that the observation ¬`high_engine_temp` does not need to be explained because it is not regarded as exceptional.

treated as 'exceptions,' namely, their negations are assumed to hold. Furthermore, for each link $\alpha \rightarrow \beta$ in the net, we include an *evidential rule* of the form $\beta \wedge \xi_i \Rightarrow \alpha$, with conjectures ξ_i with indices as indicated in figure 5.8. Finally, for those tokens γ in the network *without* incoming links, we add a rule $\gamma \Rightarrow C\gamma$; the idea is that these tokens are 'ultimate' causes and do not require explanation.

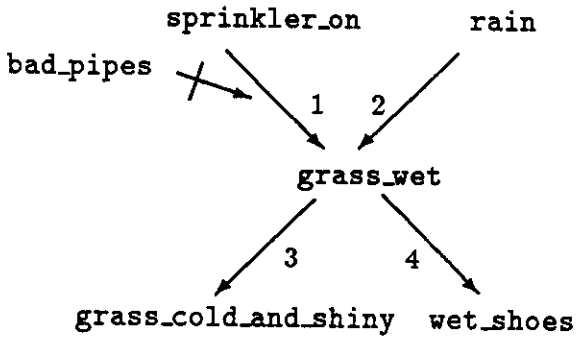


Figure 5.8: Causal and evidential defaults

Let us now assume that `grass_cold_and_shiny` is observed. The resulting context $T = \langle K, E \rangle$, with $E = \{$`grass_cold_and_shiny`$\}$ is not perfectly coherent, as the only observation has no explanation. Nonetheless, the observation 'triggers' the evidential rule `grass_cold_and_shiny` $\rightarrow$ `grass_wet`, which in turn 'triggers' the evidential rules `grass_wet` $\rightarrow$ `sprinkler_on` and `grass_wet` $\rightarrow$ `rain` That is, the evidence E gives rise to two admissible coherent belief states $\mathcal{B}_1 = \langle T, \{\xi_1, \xi_3\} \rangle$ and $\mathcal{B}_2 = \langle T, \{\xi_2, \xi_3\} \rangle$, which explain the evidence is explained in terms of `sprinkler_on` and `rain` respectively. Thus, all conjectures ξ_1, ξ_2 and ξ_3 are admissible in the context T.

If `sprinkler_on` is also observed, however, the conjecture ξ_2 associated with the evidential rule `grass_wet` $\rightarrow$ `rain` is no longer admissible. The presence of the proposition `sprinkler_on` in the new context renders the conjecture ξ_2, supporting the hypothesis `rain`, redundant. As Pearl observes, `sprinkler_on` explains `rain` away. However, if it is further learned that `sprinkler_on` is not the actual cause of `grass_cold_and_shiny` after all, say by observing `pipes_bad`, the conjecture ξ_2 becomes admissible once again, and `rain` becomes the only supported hypothesis.

It is common to find in the AI literature two different types of diagnostic tasks: *abductive diagnosis*, in which the search is for hypotheses that imply the

observations, and *consistency-based diagnosis*, in which the search is for hypotheses that render the model and the observations consistent (see [Poole, 1989]). The examples considered so far all belong to the first category. There is, however, a natural way in which consistency-based diagnosis can also be accommodated in the present framework. All that is needed is to stipulate that 'abnormalities' are self-explanatory, i.e., we need to assert expressions of the form $\alpha \Rightarrow C\alpha$ for every relevant 'abnormality' α. In that case all classes will be perfectly coherent, and thus, the causally preferred classes will be simply the *minimal classes*.

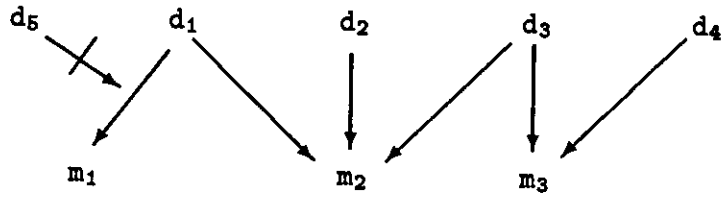
Other patterns of abductive inference, however, cannot be accommodated so easily. For instance, if the example above is supposed to reflect the situation of a farm in the Sahara dessert, we may wish to express that the hypothesis of the sprinkler being on is significantly more likely than the hypothesis of rain. The framework laid out so far, however, does not provide such facilities. For that we not only need to be able to identify the pool of conjectures, but also how such conjectures are to be ordered. The extension below addresses this limitation and shows how this additional information can be used to prune the space of admissible hypotheses.

A *preference relation on conjectures* will refer to a strict partial order on the set Ξ of conjectures. We will denote such an ordering by the symbol '$\succ$.' The expression $\xi \succ \xi'$ is to be read as stating that conjecture ξ is preferred to conjecture ξ'. Likewise, a *set of conjectures Ξ is preferred to a set of conjectures Ξ'*, if every conjecture in $\Xi - \Xi'$ is preferred to some conjecture in $\Xi' - \Xi$. Similarly, a *maximally-coherent* belief state $\mathcal{B} = \langle T, \Xi \rangle$ is a *preferred belief state* in context T, if there is no other *maximally-coherent* belief state $\mathcal{B}' = \langle T, \Xi' \rangle$ with an hypothesis set Ξ' preferred to Ξ.

By means similar to those in chapter 4, it can be shown that the preference relation among maximally-coherent belief states is also a strict partial order. Likewise, under reasonable assumptions, every preferred belief state is admissible and, furthermore, it can be identified by considering the admissible belief states only. We call the pair formed by a causal theory and an ordered set of conjectures an *abductive causal theory*. Unlike prioritized preferential structures, we will not discuss here whether there are constraints that these structures are supposed to obey. We will just illustrate their use in a simple, idealized diagnostic structure of the type considered by Reggia *et al.* [1985].

Example 5.17 Let us consider the causal network shown in fig. 5.9, We assume the $\mathtt{d}_i$'s, $i = 1, \ldots, 5$ stand for diseases, and the $\mathtt{m}_j$'s, $j = 1, \ldots, 3$ stand for manifes-

tations or symptoms. As usual, we map each link of the form $\alpha \rightarrow \beta$ into a causal rule of the form $\alpha \Rightarrow C\beta$, except the link $\mathtt{d}_1 \rightarrow \mathtt{m}_1$, which is modified by $\mathtt{d}_5$, which is expressed as $\mathtt{d}_1 \wedge \neg \mathtt{d}_5 \Rightarrow C\mathtt{m}_1$. The set $M = \{\mathtt{m}_j \mid j \in [1,3]\}$ of manifestations, determines the space of 'exceptions.' The coherence of a model will thus depend on whether the manifestations it renders true are explained.[23] We also assume that $D = \{\mathtt{d}_i \mid i \in [1,4]\}$ represents the set of possible conjectures, and that $\mathtt{d}_i$ is preferred to $\mathtt{d}_j$, $\mathtt{d}_i \succ \mathtt{d}_j$, iff $i > j$. Such preferences may be available from relevant prior probabilities.



$$\mathtt{d}_1 \succ \mathtt{d}_2 \succ \mathtt{d}_3 \succ \mathtt{d}_4 \succ \mathtt{d}_5$$

Figure 5.9: A simple diagnostic model

Let us name this background K and let us consider a context T_1 in which the manifestation $\mathtt{m}_2$ is observed. From the model depicted in fig. 5.9, it is simple to see that $\mathtt{m}_2$ gives rise to three admissible belief states $\mathcal{B}_i = \langle T_1, \{\mathtt{d}_i\}\rangle$, for $i = 1, 2, 3$. However, due to the preference order on hypotheses, only the belief state $\mathcal{B}_1$ and the hypothesis $\mathtt{d}_1$ remain preferred.

Let us further assume that, refuting the hypothesis $\mathtt{d}_1$, $\neg\mathtt{m}_1$ is observed. This new observation gives rise to three admissible belief states, but while conjectures $\mathtt{d}_2$ and $\mathtt{d}_3$ remain admissible singleton hypotheses sets, the third hypothesis set is now given by the compound hypothesis $\{\mathtt{d}_1, \mathtt{d}_5\}$. Furthermore, due to the preference order on conjectures, the hypothesis $\mathtt{d}_2$ becomes now the single preferred hypothesis, as it is preferred to $\mathtt{d}_3$ and $\mathtt{d}_5$.

Finally, let us assume that $\mathtt{m}_3$ is observed. The new context gives rise again to three admissible belief states, involving now the admissible hypothesis sets $\{\mathtt{d}_1, \mathtt{d}_4, \mathtt{d}_5\}$, $\{\mathtt{d}_2, \mathtt{d}_4\}$, and $\{\mathtt{d}_3\}$ respectively. However, due to the preferences on conjectures, this time $\mathtt{d}_3$ remains as the single leading hypothesis, followed by the

[23]The same results would follow if diseases were treated as self-explanatory exceptions as in the examples above.

compound hypotheses $\{\mathtt{d}_2, \mathtt{d}_4\}$, and only then by $\{\mathtt{d}_1, \mathtt{d}_2, \mathtt{d}_5\}$.

5.4 Related Work

Causal theories are an elaboration of ideas in [Geffner, 1989], where the notions of *explanations*, *classes*, and *coherence* were originally presented. The adoption here of a *causal operator* as part of the object-level language, however, has simplified matters considerably. This move was influenced by a proposal due to Pearl to explicitly incorporate a causal language into default theories, and by the resemblance between the preference criterion on classes advanced in [Geffner, 1989], and features of autoepistemic logic, kindly brought to my attention by Halina Przymusinska and Michael Gelfond. We consider the relation to each formulation in turn.

Pearl's proposal draws on work in causal probabilistic networks [Pearl, 1988b] to suggest that there is a natural distinction between defaults which *encode* explanations (e.g. $\mathtt{fire} \rightarrow \mathtt{smoke}$) on the one hand, and defaults which *trigger* explanations (e.g. $\mathtt{smoke} \rightarrow \mathtt{fire}$) on the other. He calls the former defaults *causal* and the latter defaults *evidential*. He argues that the language of default theories should make such a distinction clear, and in particular, that explanation 'seeking' defaults are not supposed to be triggered by explanation 'giving' defaults. Pearl's proposal to preclude such chains consist of three parts. First, he labels every default rule as either *causal*, e.g. $\mathtt{rain} \rightarrow_{\mathrm{C}} \mathtt{grass\_wet}$, or *evidential*, e.g. $\mathtt{grass\_wet} \rightarrow_{\mathrm{E}} \mathtt{sprinkler\_on}$. Secondly, he distinguishes the status of propositions p established on *causal* grounds, $\mathrm{C}p$, from those established on *evidential* grounds, $\mathrm{E}p$. Finally, he introduces a C-E calculus for reasoning with causal and evidential defaults which comprises the inference rules:

$$\frac{p \rightarrow_{\mathrm{C}} q \quad \mathrm{C}p}{\mathrm{C}q} \qquad \frac{p \rightarrow_{\mathrm{C}} q \quad \mathrm{E}p}{\mathrm{C}q} \qquad \frac{p \rightarrow_{\mathrm{E}} q \quad \mathrm{E}p}{\mathrm{E}q}$$

and which purposely excludes the rule:

$$\frac{p \rightarrow_{\mathrm{E}} q \quad \mathrm{C}p}{\mathrm{E}q}$$

Though differing in detail and goals, the appeal to a causal operator in the

context of causal theories is based on similar intuitions. The reading, though not the interpretation of the causal expression Cp, is indeed the same in Pearl's C-E system as in causal theories. On the other hand, the distinction between causal and evidential default rules is captured here in terms of the distinction between *assumptions* (defaults) and *conjectures*. Conjectures, unlike assumptions, are triggered *only* by the need to explain "abnormalities." A similar approach has been indeed previously advanced by Poole [1987].

The relation between causal theories and autoepistemic theories (see [Moore, 1985b], and section 1.5) can be best illustrated in the domain of general logic programs. A propositional rule of the form

$$\gamma \leftarrow \alpha_1, \ldots, \alpha_n, \neg\beta_1, \ldots, \neg\beta_m \quad ,$$

in a program P, is translated as a rule:

$$\alpha_1 \wedge \ldots \wedge \alpha_n \wedge \neg \mathrm{L}\beta_1 \wedge \ldots \wedge \neg \mathrm{L}\beta_m \Rightarrow \gamma$$

in the autoepistemic theory $L[P]$ [Gelfond, 1987], and as a causal rule:

$$\mathrm{C}\alpha_1 \wedge \ldots \wedge \mathrm{C}\alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow \mathrm{C}\gamma$$

in the causal theory $C_1[P]$ which, provided P is stratified, *legitimize the same (plain) literals*. Indeed, if we look at the *forms* of the autoepistemic theory $L[P]$ and the causal theory $C_1[P]$, it is possible to appreciate that the operators L and C behave as *duals* in some sense. Namely, we can go from one encoding to the other by *removing* either autoepistemic or causal operators from certain atoms, and by *adding* either causal or autoepistemic operators to every other atom, while preserving the same meaning of the theory.

Actually it is possible to understand the autoepistemic operator L as an *evidential* operator, with Lα meaning: there is evidence for α. Instead of using the *causal* operator C under the conventions that

$\neg\alpha$ is an assumption
C$\alpha \Rightarrow \alpha$ must hold for every (plain) sentence α, and
α is *explained* in a class when Cα holds,

we could have instead used an *evidential* operator E under the conventions that

$\neg E\alpha$ is an assumption
$\alpha \Rightarrow E\alpha$ must hold for every (plain) sentence α, and
$E\alpha$ is *explained* in a class when α holds.

Under such an approach the *evidential* encoding of a logic program would be *identical* to the autoepistemic encoding, except for the presence E's instead of L's. Moreover, both encodings would sanction an equivalent semantics for stratified programs.

For non-stratified programs, however, as for most default theories, the duality between causal and autoepistemic disappears. First, default theories may lack stable models.[24] More importantly, the behavior of an autoepistemic rule $\neg Lp \Rightarrow q$ departs from the behavior of a causal $\neg p \Rightarrow Cq$, when $\neg p$ is not an *assumption*. Namely, autoepistemic theories, unlike causal theories, do not require a an identified set of assumptions in the language; the prefix $\neg$L, as no causal prefix, generates the assumptions needed.

There has also been a lot of work in philosophy around the notions conditionals, causes and explanations.[25] For the most part, however, philosophers have been concerned with making the meaning of such notions precise, while we have been mainly concerned with exploiting the basic intuitions for designing better knowledge representation languages and semantics. Both goals, however, are highly inter-related, and a closer interaction between them is to be expected.

[24]Though see the extension of autoepistemic logic due to Gelfond and Przymusinska [1989].

[25]See for instance, the collection of papers in [Pitts, 1988] and the recent book by Gardenfors [1988].

Chapter 6

Conclusions

The construction and analysis of programs capable displaying the reasoning abilities of people requires the representation of rich fragments of commonsense knowledge. Representation languages suited for this task must provide expressive and meaningful primitives in which knowledge can be naturally encoded and efficiently processed. In this work we have addressed issues relevant to the design of languages for representing defeasible knowledge. In particular, we have been concerned with the form and interpretation of default theories.

6.1 A New Interpretation of Defaults

Defaults play a central role in commonsense reasoning, permitting the generation of useful predictions in the absence of complete information. Nonetheless, the fact that these predictions are non-monotonic has made an understanding of default reasoning in terms of the traditional tools of formal logic difficult. Formal interpretations of default reasoning had thus appealed to non-monotonic extensions of classical logic. However, in recent years, work in inheritance hierarchies, temporal reasoning, logic programming and abductive reasoning has pointed to aspects other than non-monotonicity that must also be addressed when reasoning with defaults.

The thesis advanced in this work is that most of these aspects have to do with two dimensions of default reasoning often ignored by non-monotonic logics: a conditional dimension, by which default expectations are regarded as conditional

assertions; and a causal dimension, by which explained expectation failures are distinguished from unexplained ones. The conditional interpretation of defaults evolved from a probabilistic interpretation into a form of entailment, called *conditional entailment,* which combines conditionals and defaults. The model and proof theories of conditional entailment take a form similar to those advanced for prioritized circumscription. However, while priorities in circumscription are a means for the user to express preferences among assumptions, priorities in conditional entailment are extracted automatically from the knowledge base. Such priorities enforce the conditional reading of defaults; namely, they permit us to assert a proposition q from a body of *evidence* $\{p\}$ in a *background context* containing a default $p \rightarrow q$, regardless of the presence of conflicting defaults. The distinction between a background context conveying generic information and a body of evidence conveying factual information is critical for such an interpretation to work.

Causal aspects are captured by extending the language of default theories by means of a causal operator. Such an operator is used to express and identify the conditions under which abnormalities are regarded as "explained." Classes of models which succeed in explaining the abnormalities they engender are then rewarded, determining the propositions which are causally entailed. Causal entailment suffices to account for theories which lack a conditional component such as general logic programs. Domains such as inheritance hierarchies, on the other hand, require both causal and conditional considerations to be taken into account. In that case, the entailment relation is defined in terms of the causally preferred *conditional* classes of the theory in question.

The main contribution of this research is to formulate and point to two aspects of defaults which suffice to provide a reasonable account of a variety of domains of interest in AI, including inheritance hierarchies, reasoning about change, general logic programs and abductive reasoning. With the exception of abductive reasoning, all these domains accept a natural and faithful representation in terms of causal theories. Abductive theories, on the other hand, require an additional component: a distinguished and possibly ordered set of *conjectures*. Admissible sets of conjectures represent sets of hypotheses which an agent may find reasonable to postulate in order to make sense of a given body of evidence. Together with the information available, the admissible sets of conjectures represent the *belief states* to which an agent is likely to commit.

6.2 Loose Ends

Having summarized the good news, it is time now to recount aspects which have not received a satisfactory treatment. Two such aspects will elaborated in some detail: the first is regarding the integration of causal and conditional considerations proposed in section 5.2.3, the second, regarding languages, architectures and *real* reasoning as opposed to default reasoning.

Causal and conditional preferences

The proposed integration of causal and conditional preferences consists of two steps. First, the preferred models of the theory T under consideration are grouped into what we call the conditional classes of T, and then the causally preferred classes among them are selected. The propositions entailed by T are then identified as those that hold in such classes.

This two-step process is not completely satisfactory, and it hinders the development of a concise proof-theory for the resulting entailment relation. A cleaner integration, more amenable to analysis, would require merging causal and conditional preferences into a unique priority ordering on assumptions. However, while we have identified a family of admissible priority orderings which reflect conditional preferences on assumptions, nothing of that sort has been achieved for causal preferences, which are exclusively determined by the existence of explanations. In the following paragraphs we will discuss issues relevant to a 'prioritized' understanding of causal preferences, and possible ways to integrate causal and conditional priorities.

Let us recall first the relevant notions. First, the negation of an assumption δ is explained in a class $\mathcal{C}$ of a causal theory T when the literal $\mathrm{C}\neg\delta$ holds in $\mathcal{C}$. The set of assumptions in the gap $\Delta[\mathcal{C}]$ of $\mathcal{C}$, whose negations are explained constitute the explained gap of $\mathcal{C}$. Gaps and explained gaps determine the causal ordering on classes. Namely, the class $\mathcal{C}$ is as preferred as a class $\mathcal{C}'$ iff every assumption in $\Delta[\mathcal{C}] - \Delta[\mathcal{C}']$ belongs to the *explained* gap of $\mathcal{C}$, and is preferred to $\mathcal{C}'$ iff it is as preferred as $\mathcal{C}'$ but $\mathcal{C}'$ is not as preferred as $\mathcal{C}$.

A possible way to understand such preference on classes in terms of assumption priorities, is to appeal to the proof-theoretic conditions under which exceptions are explained. As we said in section 5.2.2, the literal $\mathrm{C}\neg\delta$ holds in a class $\mathcal{C}$ in a context T, when there is a set of assumptions Δ outside the gap of $\mathcal{C}$, that together with

T logically imply C$\neg\delta$. Let us assume, furthermore, that Δ is a *minimal* such set, and let us call it a *minimal causal support* of $\neg\delta$ in T. Then, it must be the case that for every assumption δ in $\Delta[\mathcal{C}] - \Delta[\mathcal{C}']$ which belongs to the explained gap of $\mathcal{C}$, there is an assumption δ' in $\Delta[\mathcal{C}']$ which participates in a minimal causal support of $\neg\delta$. Thus, if we establish a non-strict (i.e., reflexive and transitive) preference relation '$\preceq_C$' on assumptions such that $\delta \preceq_C \delta'$ holds in a context T, if δ' belongs to a minimal causal support of $\neg\delta$ in T, a causal preference relation '$<_C$' on classes can be obtained as follows.

A class $\mathcal{C}$ would be *as preferred as* a class $\mathcal{C}'$ in T, written $\mathcal{C} \leq_C \mathcal{C}'$, if for every assumption δ in $\Delta[\mathcal{C}] - \Delta[\mathcal{C}']$ there exist an assumption δ' in $\Delta[\mathcal{C}'] - \Delta[\mathcal{C}]$ such that $\delta \preceq_C \delta'$. Similarly, a class $\mathcal{C}$ would be *preferred to* a class $\mathcal{C}'$, written $\mathcal{C} <_C \mathcal{C}'$, when $\mathcal{C} \leq_C \mathcal{C}'$ holds, but $\mathcal{C}' \leq_C \mathcal{C}$ doesn't.

The entailment relation that follows from this type of preferences is not equivalent to causal entailment, though it preserves its main features. The departure is a consequence of the fact that preferences based on explanations — which are between *sets* of assumptions, on the one hand, and single assumptions on the other — have new been projected onto a a preference relation '$\preceq_C$' on *pairs* of assumptions. Likewise, we have made such a preference relation transitive. Nonetheless, the behavior rendered by both causal accounts is equivalent within a sufficiently large family of causal theories as to regard the new account as a tentative substitute. Its main appeal is that it permits to map the problem of combining conditional and causal preferences into the apparently more manageable problem of combining the conditional priority orderings '$\prec$' determined by a background K, with the non-strict causal preferences '$\preceq_C$' dictated by a particular context $T = \langle K, E \rangle$. There are many ways in which to proceed with such a combination. We will not go into these details here. Rather we will mention some of the features that would make one such combination adequate.

First, a preference on assumptions based on causal and conditional preferences has be to such that its effect is similar to the two-step process described above; namely, the ultimate preferred classes of a causal theory T must be among the conditional preferred classes of T. The reason is that explanations in non-conditionally preferred classes are often spurious, and cannot be relied upon for coherence judgements. Second, the resulting preference relation on assumption has to render a behavior equivalent to conditional entailment in those domains which lack a conditional component, and in which causal entailment behaves properly (e.g. causal theories associated with stratified programs). Finally, it has to be simple and amenable to a proof-theory similar to that of conditional entailment. In summary, it has to capture the behavior associated with the causally preferred

conditional classes of a theory, without having to rely on such a crisp distinction between causal and conditional considerations.

Languages, Architectures and Reasoning

We have been talking all along about default reasoning. It would be good now to stop for a while and wonder about what is the relation between *real* reasoning and default reasoning in the sense we have assumed. Real reasoning is what intelligent people do to accomplish their goals; default reasoning, on the other hand, is the name for an abstract family of patterns of inference whose main characteristic is its reliance on rules which admit exceptions. There is nothing like *real* default reasoning; real reasoning exhibit traits of deductive, default, abductive and other forms of reasoning all combined. Moreover, while the analysis of these different reasoning forms is concerned with describing inferences which are likely to yield true conclusions from true premises, real reasoning is concerned with inferences which are likely to be useful. While inferences leading from true premises to false conclusions are likely not to be useful, the likelihood of arriving to true conclusions from true premises is just one of the dimensions a reasoning agent must be concerned with. The other dimension is regarding the resources available; a clever chess program designed for a CRAY supercomputer is likely not to fare well against a clever chess program designed for a PC when both programs are run on a PC.

With this perspective in mind, it is still possible for formal analyses of abstract forms of reasoning to contribute to a better understanding of real reasoning, and to the construction of programs capable of displaying some of the reasoning abilities of people. In this last regard, such formal analyses should not only provide the vocabulary on which research progress on the topic could be articulated, but also useful guidelines for the design of such programs. In the area of our concern — default reasoning— some of the questions that need to be addressed are: (1) what is a good *language* for expressing default knowledge, (2) what is a good *architecture* for reasoning with default knowledge, and (3) what is a 'reasonable' default *reasoning task.*

In this work we have partially addressed questions (1) and (3), but only in an isolated form. At the end, however, such question require an integrated answer: the representation language of programs capable of reasoning with defaults must be designed with a reasoning architecture in mind, and such architecture will

be heavily influenced by the reasoning task envisaged. The work reported here attempts to provide a better understanding of what default reasoning is, and what the requirements for a language which supports defaults reasoning are (e.g. K vs. E). Now we would like to know more about what the proposed framework reveals about *default reasoning architectures*. Before expressing some ideas in that regard, however, we will find useful to review some of the successful *language-architecture-reasoning task* triplets existent in the knowledge representation arena. We will refer to them as *reasoning boxes*.

One such reasoning box is *inheritance reasoning*. Inheritance networks are both a language and an architecture which support a particular reasoning task called inheritance reasoning (see section 1.4). Inheritance reasoning is concerned with finding "good" inheritance paths in inheritance hierarchies. The limited expressive power of inheritance networks and the narrow focus of inheritance reasoning enable efficient reasoning, hard to achieve in more expressive frameworks.[1] Another appealing reasoning box based on networks, is *Bayesian Networks* [Pearl, 1988b]. Like inheritance networks, Bayesian Networks are both a language, in this case to express probabilistic knowledge, and an architecture. Several reasoning tasks have been defined on those networks, including the computation of degrees of beliefs and most likely belief commitments. Furthermore, Pearl has shown that such tasks can be performed in a highly efficient manner, for Bayesian networks complying with certain constraints.

Other successful reasoning boxes are logic programs and truth maintenance systems. The language of general logic programs is that of universally quantified rules, and their standard architecture is that of SLD resolution, augmented by negation as failure [Lloyd, 1984]. The task in the domain of logic programs can be understood, in essence, as computing the atomic consequences that follow from a logical interpretation of the program in question (e.g. the completion of the program [Clark, 1978]).[2] Moreover, SLDNF resolution is a sound and complete method for performing such task within a large class of programs [Lloyd, 1984]. It is interesting to note that logic programs do not fall into the "irrelevance" trap of pure logic; namely, while $p \vee p$ and $p \wedge p$ and infinitely many other formulas are logical consequences of p, SLDNF, for instance, never bothers to prove them. For the reasoning task adopted, they are *known* to be irrelevant.

Finally, assumption truth maintenance systems (ATMSs) are systems designed

[1] Though see the worst-case results investigated by Selman and Levesque [1989].

[2] Logic programs also compute answer substitutions. Still, these substitution are computed in terms of the atomic consequences of the program.

for keeping track of results which rely on assumptions which may later have to be retracted. They evolved from early work by Stallman and Sussman [1977] on ruled-based programs for analyzing circuits, and Doyle's [1979] TMS. de Kleer [1986] presented an informal description of the standard ATMS, as of its language and architecture. The ATMS task has been later described in detail by Reiter and de Kleer [1987]. Roughly, the ATMS is given a set T of propositional facts and rules embedding assumptions, and computes the minimal sets of assumptions Δ consistent with T, which together with T logically imply particular literals of interest. This computation is usually done by keeping track of the minimal set of assumptions logically inconsistent with T (no-goods). Still, since the task is highly intractable in the case of non-Horn rules, the ATMS sometimes only approximates such a behavior, being unsound and incomplete at times.

These successful reasoning boxes, raise the question of whether a suitable reasoning box for default reasoning can be defined. The answer, I think, is a qualified yes. For conditional entailment, for instance, we may want to consider 'almost' propositional theories; namely, universal rules and atomic queries on first order languages with a finite number of constants and no function symbols. Furthermore, it makes sense to restrict attention to background contexts which accept a single *minimal* admissible priority ordering '$\prec$'. As shown in section 4.2, this will permit us to compute conditional entailment by considering a single prioritized admissible structure $\langle \mathcal{I}_{\mathcal{L}}, <, \Delta_{\mathcal{L}}, \prec \rangle$. The reasoning task and architecture will bear some resemblance to a *prioritized ATMS*. Namely, rather than computing minimal supports Δ for particular literals p, such that Δ is *consistent* with the theory theory $T = \langle K, E \rangle$ in question, we would compute minimal supports Δ for p, such that Δ is *stable* in the context T (see definitions in section 4.3). Such literals p would then represent propositions conditionally entailed by T. Not every conditionally entailed proposition would be identified, though: first, we would be focusing on literals rather than on arbitrary formulas; second, a literal may be conditionally entailed and still lack a stable support (section 4.3). Nonetheless, the claim is that literals supported by stable sets will represent the main propositions of interest. Note also that the switch from *consistency* in the ATMS to *stability* in a prioritized ATMS, is a result of the change from the *minimal model semantics* underlying the ATMS task (see [Ginsberg, 1989], for instance) to the *prioritized minimal model semantics* underlying conditional entailment.

The details of such a prioritized ATMS will be elaborated elsewhere. The main intuition, however, is simple. By the results in section 4.3, in order to test whether the set of assumptions Δ is stable, we can first identify all the minimal set of assumptions Δ' from which Δ is *protected*. that is, we look in the ATMS database

for no-goods of the form $\Delta' + \Delta_s$, where Δ_s is a subset of Δ such that $\Delta' \prec \Delta_p$ holds. The set Δ then is stable iff every no-good Δ'' which contains an assumption from Δ, is such that $\Delta'' - \Delta$ is one of the sets Δ' from which Δ is protected.

An architecture of that sort for capturing causal entailment as well, will require the development of an appropriate proof-theory for causal entailment. It is also to expect that the explicit use of the causal operator 'C' could be avoided by appealing to a suitable *network* languages, such as the language of causal networks described in sections 5.3.3 and 5.3.4.

Finally, the search for algorithms should not be exclusively focused on those which are complete or even sound with respect to some accepted formalization. Tractable algorithms which are 'reasonably' sound and 'reasonably' complete, and whose sources of unsoundness and incompleteness can be understood and justified, are likely the best to be achieved in default reasoning and reasoning in general.[3]

6.3 Open Problems

In this final section we will discuss some relevant open problems. The focus is on the *projection problem* in reasoning about change and belief revision, and the interactions between probabilistic and non-monotonic reasoning.

The Projection Problem

Consider the following story from [Maloney, 1989] in relation to the frame problem:

> Eloise glances at her window and sees Abelard standing in the courtyard, wearing a hat, leaning against the chestnut tree and speaking to her father ... Soon one of her brothers enters and tells her that Abelard has departed. Now Eloise has to reasses her beliefs. Minimally, she must delete her belief that Abelard is in the courtyard. But what of his hat? ...

Intuitively, the belief that Abelard's hat is in the courtyard should be deleted as well. However, the hat *was* in the courtyard, and there is no explicit information in

[3]See for instance the study of the complexity of different fragments of Reiter's defaults logic in [Kautz and Selman, 1989].

conflict with the assumption that the hat remained in the courtyard. Thus, if the location of the hat is regarded as a fluent (namely, a time or situation dependent property [McCarthy and Hayes, 1969, McDermott, 1982]), and the persistence of fluents is assumed by default, we would be led to the counterintuitive conclusion that the hat remained in the courtyard. On the other hand, if we do not regard the location of the hat as a persistent fluent, we wouldn't be able to predict the location of the hat after having dropped it in the courtyard.

The lesson to be drawn from this and similar examples is that the assumption that all fluents persist in time by default, while a useful one to make, is not generally tenable. In most contexts there will be some fluents whose persistence can be assumed to hold in the lack of contrary evidence, any many other fluents which cannot.

How to handle the persistence of fluents then? Let us refer to the proposition f_t denoting the status of a fluent f at time t, in some appropriate notation, as a *propositional fluent*, and let us say that a propositional fluent f_t is *projectible* when it is reasonable to jump to the proposition f_{t+1} by default. In the story above, for instance, when Abelard is in the courtyard, the location of Abelard is projectible, while the location of the hat is not. The *projection problem* — to be distinguished from the more general problem of temporal projection or prediction (e.g. [Hanks and McDermott, 1987]) — is the problem of identifying the projectible propositional fluents in a given context.

Several proposals have been recently advanced for addressing this and related problems. These proposals range from those which presume that the projectible propositional fluents have been identified by the user (e.g. [McDermott, 1982, Kowalski and Sergot, 1986]), to those in which persistence defaults contain appropriate censors (e.g. [Myers and Smith, 1988]). Likewise, in some approaches the projectibility of a propositional fluent is determined by their (explicit) appearance in the context in question (e.g. [Ginsberg and Smith, 1988]), while in others, all propositional fluents are assumed equally projectible (e.g. [Winslett, 1988]). However, while these proposals contain elements which are necessary in dealing with the projection problem; none addresses it in its full generality. Myers and Smith for instance, who are among the few to recognize the problem as such, regard it as the problem of determining the persistence of *derived information*. Namely, they assume that what we call the projectibility of a propositional fluent depends on its *derivation* from the information explicitly available.

In my view, the projection problem is more general and concerns the identification of the *independent* propositional fluents in a particular context. For instance,

the propositional fluent $\mathtt{above(a,b)}_t$ is projectible in a context which does not contain other information, but is not projectible in a context which also includes the propositional fluents $\mathtt{on(a,c)}_t$ and $\mathtt{on(c,b)}_t$. Similarly, the propositional fluent $\neg\mathtt{on(a,table)}_t$ is projectible if alone, but is not projectible when augmented with an additional propositional fluent $\mathtt{on(a,c)}_t$. In each case, we need to conjecture what *depends* on what and project only what we regard as *independent*. Like in abductive reasoning (section 5.3.4), however, there may be several equally good conjectures in a particular context. The projection problem is the problem of elucidating the nature and the logic of these conjectures.

Non-monotonic Reasoning and Probabilistic Reasoning

Another area open for exploration lies on the boundary between non-monotonic and probabilistic reasoning. Several of the notions developed in this work, such as the semantics for the core and the account of abduction, rely on a probabilistic basis. It would be interesting to investigate if there are other probabilistic notions can be imported in a qualitative framework of defeasible inference. Similarly, the question arises whether the qualitative frameworks of defeasible inference may be used to extend probabilistic reasoning with assumptions about conditional independence. This topic has been explored by Grosof [1988].

Grosof focuses on the conditions under which the value of an entry $P(H \mid E)$ of a partially specified probability distribution P can be approximated, given that the only known entries $P(H \mid C_i)$, $i = 1, \ldots, n$ are such that the propositions C_i constitute a chain; namely, $P(C_{i+1} \mid C_i) = 1$, for $i = 1, \ldots, n-1$. On the basis of an intuition similar to Kyburg's [1983] notion of reference classes, Grosof assumes that the value of $P(H \mid E)$ can be approximated to $P(H \mid C_i)$, when C_i is the first member of the chain for which $P(C_i \mid E) = 1$. Furthermore, using Nilsson's [1986] probabilistic logic, he shows how such form of non-monotonic reasoning about probabilities can be formulated in McCarthy's circumscriptive framework. In the next few lines we show how conditional interpretations of defaults, with their ability to capture the context-dependent nature of conditional probabilistic statements might provide an alternative framework in which to reason about probabilities.

When reasoning about probabilities, we are interested in determining the probability P of an hypothesis H given certain constraints C on P, and a body of evidence E. We will cast this problem as the problem of encoding the constraints C on P in a background context K, in such a way that if $P(H|E) = x$ is the target result, the proposition $B(H) = x$ is conditional entailed by $T = \langle K, E \rangle$.

For simplicity we will focus on a variable-free first order language with equality, with an operator $P(\cdot)$ which complies with the axioms of probability, and an indexical belief operator $B(\cdot)$ such that in context $T = \langle K, E \rangle$, $B(H)$ represents the probability statement $P(H \mid E)$.

The direct encoding in K of a probability statement $P(q|p) = x$ is as a default $p \rightarrow [B(q) = x]$. Then, given a body of evidence $E = \{p\}$, the proposition $B(q) = x$ will be conditionally entailed, correctly assessing that the belief in q given a body of evidence $E = \{p\}$ is equal to $P(q \mid p)$. More interestingly, when the body of evidence E is extended with an additional, but irrelevant piece of information e, the conclusion $B(q) = P(q \mid p)$ (hence, the belief in q) remains unaltered. Grosof's idea of default inheritance of probabilities, follows as a natural consequence of such an encoding, provided certain conditions are met. First we need an axiom in K of the form $p \Leftrightarrow [B(p) = 1]$, to relate p and the belief operator B. Then, given a set of defaults $p_i \rightarrow [B(\mathrm{H}) = P(\mathrm{H} \mid p_i)]$, $i = 1, \ldots, n$, such that the propositions $p_1, p_2, \ldots, p_n$ form a chain, as above, the proposition $B(\mathrm{H}) = P(\mathrm{H} \mid p_i)$ will be conditionally entailed by the theory $T = \langle K, E \rangle$, as long as p_i is conditionally entailed by E, and E does not provide support to any of the propositions p_j in the chain for which $j < i$ (see the irrelevance rule in section 2.5).

Clearly, these remarks only scratch the surface of the problems involved in the use of conditionals for reasoning about probabilities. For instance, the encoding suggested above cannot deal with *probabilistic chains*; namely two probabilistic statements $P(q \mid p) = x$ and $P(p \mid r) = y$, do not yield a judgement regarding the value of $P(q \mid r)$. Similarly, the encoding of two probabilistic statements $P(q \mid p) = x$ and $P(q \mid r) = y$ for $x \neq y$, yields that $P(q \mid q, r)$ is either $P(q \mid p)$ or $P(q \mid r)$, contrary to the intuition suggesting that the supports of p and r should be somehow combined. Whether it is possible to capture commonsense probabilistic arguments in a default framework is thus to be seen. An effort in that direction may also help understand the relations between probabilistic inference, fuzzy logic [Yager *et al.*, 1987] and Dempster-Shafer inference [Shafer, 1976], which also appear to rely heavily on assumptions about independence.

Appendix A

Proofs

Theorem 2.2 *If $E \mathrel{|\!\sim}_K p$ and $K \subseteq K'$ then $E \mathrel{|\!\sim}_{K'} p$*

Proof The theorem easily follows by induction on the minimal length n of the derivation of $E \mathrel{|\!\sim}_K p$. If $n = 1$, it means that h was derived from E in K either by rule 1 or by rule 2. In either case it is easy to show that h can be derived from E in K'. Let us assume now that h is derivable from E in K in n steps, $n > 1$, and that the theorem holds for all the proofs with length $m < n$. Clearly the last step in the derivation must involve one of the rules 3–5. In any case, the antecedents of such a rule must be derivable in a number of steps smaller than n and, therefore, by the inductive assumption, they are also derivable in K', from which it follows that, using the same rule, h is also derivable from E in K'.

Theorem 2.3 *If $E \mathrel{|\!\sim}_K p$, then p is ϵ-entailed by the default theory $T = \langle K, E \rangle$.*

Proof We show this by proving each rule in the core to be sound with respect to ϵ-entailment. That is, for a rule with conclusion $E \mathrel{|\!\sim}_K p$, we prove that for any probability distribution P admissible with K within a range ϵ and which complies with the premises of the rule, the probability $P(p \mid E)$ must approach one, as ϵ approaches zero. The **defaults** rule is sound by definition: if $p \rightarrow q$ is a default in K, then the admissibility of P requires the conditional probability of $P(p \mid q)$ to approach one as ϵ approaches zero. For **deduction**, if $E \mathrel{|\!\sim}_K p$, then, clearly $P(p \mid E) = 1$. To prove the soundness of **augmentation** and **reduction**, we need to show that if $P(p \mid E)$ approaches one, $P(q \mid E, p)$ approaches one iff $P(q|E)$ does. This, in turn, is a consequence of the probabilistic equality:

$$P(q \mid E) \;=\; P(q \mid E,p)\; P(p \mid E) \;+\; P(q \mid E,\neg p)\; P(\neg p \mid E)$$

Indeed, if $P(q \mid E)$ and $P(p \mid E)$ approach one, so must $P(q \mid E,p)$, since $P(\neg p \mid E)$ approaches zero, and $P(q \mid E,\neg p)$ is bounded by one. On the other hand, since the value of $P(q \mid E)$ is bound from below by the product of $P(q \mid E,p)$ and $P(p \mid E)$, the former term must approach one, as the latter two terms do. Finally, we prove the soundness of **disjunction** by proving the soundness of **weak reduction** first, which permits to derive $E \mathrel{|\!\sim_K} \neg p \vee r$ from $E, p \mathrel{|\!\sim_K} r$. The soundness of the rule is obvious in case $E \not\mathrel{|\!\sim_K} \neg p$ holds. Otherwise, by Bayes rule we have:

$$P(\neg r \wedge p \mid E) \;=\; P(\neg r \mid E,p)\; P(p \mid E)$$

In particular, if $P(r \mid E,p)$ approaches one, $P(\neg r \mid E,p)$ must approach zero, and so does $P(\neg r \wedge p \mid E)$. As a result $P(\neg(\neg r \wedge p) \mid E)$ must approach one, and so $P(r \vee \neg p \mid E)$, due to the logical equivalence between $\neg(\neg r \wedge p)$ and $r \vee \neg p$. To derive now the disjunction rule, note that the augmentation rule permits us to derive $E, p, p \vee q \mathrel{|\!\sim_K} r$ and $E, q, p \vee q \mathrel{|\!\sim_K} r$, from $E, p \mathrel{|\!\sim_K} r$ and $E, q \mathrel{|\!\sim_K} r$. Weak reduction then permits us to obtain $E, p \vee q \mathrel{|\!\sim_K} r \vee \neg p$ and $E, p \vee q \mathrel{|\!\sim_K} r \vee \neg q$. Finally, from these two expression we can obtain $E, p \vee q \mathrel{|\!\sim_K} r$ by deductive closure. ∎

Theorem 2.4 *The proposition q is ϵ-entailed by the default theory $T = \langle K, E\rangle$, with $K = \langle L, D\rangle$ and $E = \{p\}$, if and only if $K' = \langle L, D + \{p \rightarrow \neg q\}\rangle$ is ϵ-inconsistent.*

Proof The only-if part of the theorem is a simple consequence of the monotonicity of the relation '$\mathrel{|\!\sim_K}$' (theorem 2.2) with respect to K. If q follows from p in K, q will certainly follows from p in K'. However, since $\neg q$ also follows from p in K', due to the soundness of rules 1–5, K' cannot be ϵ-consistent. The other half of the theorem is simple prove in the case in which K itself is ϵ-inconsistent. In this case E ϵ-entails any sentence in the language. We will assume that K is ϵ-consistent, and show that p ϵ-entails q in K when K' is ϵ-inconsistent. We will follow the proof in Adams [1975], also sketched in [Goldszmidt and Pearl, 1989], and rely on results to be fully established in chapter 3. Two useful concepts in the proof are the notions of default verification and falsification, and quasi-conjunctions. A default $p \rightarrow q$ is *verified* in a world W (i.e. a truth valuation) iff W satisfies both p and q, and is *falsified* in W iff W satisfies p but not q. Likewise, provided there is finite set of worlds which satisfy the sentences in L, a background context

$K' = \langle L, D' \rangle$ is ϵ-inconsistent, only D' contains a set D'', such that every world that satisfies L and *verifies* a default in D'', must also *falsify* a default in D'' (first part lemma 3.4, and lemma 3.5, chapter 3). Under the assumptions above, this implies that if $K' = \langle L, D + \{p \rightarrow \neg q\} \rangle$ is ϵ-inconsistent, there must be one such set D'' which must contain the default $p \rightarrow \neg q$.

The *quasi-conjunction* $C(D)$ of a set D of defaults $p_i \rightarrow q_i$, $i = 1, \ldots, n$, is the default $C(D) = p_1 \vee p_2 \vee \cdots \vee p_n \rightarrow (p_1 \Rightarrow q_1) \wedge (p_2 \Rightarrow q_2) \wedge \cdots \wedge (p_n \Rightarrow q_n)$. Due to the logical equivalence between the formulas $\neg(p_1 \Rightarrow q_1) \vee \cdots \vee \neg(p_n \Rightarrow q_n)$ and $\neg(p_1 \vee p_2 \vee \cdots \vee p_n \Rightarrow (p_1 \Rightarrow q_1) \wedge (p_2 \Rightarrow q_2) \wedge \cdots \wedge (p_n \Rightarrow q_n))$ the existence of a set D'' in K' as above, implies that there is no world that satisfies L and verifies the quasi-conjunction $C(D'')$ of D''. Now, let $U_P(D_0)$ stand for the sum:

$$U_P(D_0) = \sum_{i=1}^{n} 1 - P(q_i|p_i) \quad ,$$

for a probability distribution P and a set D_0 of defaults $p_i \rightarrow q_i$, $i = 1, \ldots, n$, and consider the sums $U_P(D_0)$ and $U_P(\{C(D_0)\})$, where $C(D_0)$ is the quasi-conjunction of D_0. $U_P(\{C(D_0)\})$ represents the sum of the probabilities over all the worlds that falsify a default in D_0, while $U_P(D_0)$ includes all such terms, and possibly, many more. $U_P(D_0)$ is thus greater than $U_P(\{C(D_0)\})$ for any P and set D_0. In particular, $U_P(D'') \geq U_P(\{C(D'')\})$, and in the context of a probability P which assigns unit probability to the sentences in L, $U_P(D'') \geq 1$, since as we showed, no world satisfies L and verifies $C(D'')$ and thus $U_P(\{C(D'')\}) = 1$. Moreover, if P is a probability distribution admissible with K within a range ϵ, as ϵ approaches zero all terms in $U_P(D'')$ which correspond to defaults in D will vanish. In the limit, since $p \rightarrow \neg q$ is the only default in D'' which is not in D, we get that $1 - P(\neg q \mid p)$ and thus $P(q \mid p)$ must approach one. Thus, q is ϵ-entailed by p in K. ∎

Theorem 3.1 *If the expression $E \mathrel{\tilde{\vdash}}_K p$ is interpreted as asserting that p is true in all preferred models of the default theory $T = (K, E)$ of every preferential model structures well-founded with respect to K, then the following rules are sound:*

Rule 2 (Deduction) If $E \vdash_K p$ then $E \mathrel{\tilde{\vdash}}_K p$
Rule 3 (Augmentation) If $E \mathrel{\tilde{\vdash}}_K p$ and $E \mathrel{\tilde{\vdash}}_K q$ then $E, p \mathrel{\tilde{\vdash}}_K q$
Rule 4 (Reduction) If $E \mathrel{\tilde{\vdash}}_K p$ and $E, p \mathrel{\tilde{\vdash}}_K q$ then $E \mathrel{\tilde{\vdash}}_K q$
Rule 5 (Disjunction) If $E, p \mathrel{\tilde{\vdash}}_K r$ and $E, q \mathrel{\tilde{\vdash}}_K r$ then $E, p \vee q \mathrel{\tilde{\vdash}}_K r$

Proof **Deduction** is clearly sound: if p is true in all the models of E, p will be true in all the preferred models of E. The soundness of **augmentation** and **reduction** follows from the fact that in any well-founded structure $\langle \mathcal{I}, < \rangle$ where p is true in all the preferred models of E, the preferred models of E and $E \cup \{p\}$ coincide. Assume otherwise, that there is a well-founded p-structure in which M is a preferred model of E but not of $E' = E \cup \{p\}$. Clearly, M must a be *model* of E' since p holds in all preferred models of E. Then, by well-foundness, there must be a preferred model of M' of E' such that $M' < M$. However, since M' is also a model of E, that would contradict the assumption that M is a preferred model of E. A similar contradiction results if we assume that there is an interpretation N which is a preferred model of $E' = E \cup \{p\}$ but not of E. That would imply that there must be a preferred model N' of E, such that $N' < N$. That, however, would contradict N being a preferred model of E', as E' is also satisfied by N'. Finally, **disjunction** follows from the fact that in any p-structure, the preferred models of a disjunction $\alpha \vee \beta$ are among the preferred models of α and the preferred models of β. Assume otherwise a well-founded p-structure $\langle \mathcal{I}, < \rangle$ with a model M which is preferred for $\alpha \vee \beta$ but not for either α or β. Clearly, M must be a model of either α or β. Without loss of generality we can assume M to be a model of α. By well-foundness then, there must be a preferred model M' of α such that $M' < M$. Since M' is also a model of $\alpha \vee \beta$, however, this implies that M cannot be a *preferred* model of $\alpha \vee \beta$, in contradiction with the former assumption. ∎

Theorem 3.2 *If the proposition p is is derivable from a context $T = \langle K, E \rangle$ by means of rules 1–5, then p is preferentially entailed by $T = \langle K, E \rangle$.*

Proof Straightforward from theorem 3.1 and the definition of p-entailment. ∎

Lemma 3.1 *A default theory $T = \langle K, \{p\} \rangle$ with a background context $K = \langle L, D \rangle$ p-entails a sentence q if only if the background $K' = \langle L, D \cup \{p \rightarrow \neg q\} \rangle$ is p-inconsistent.*

Proof The 'only if' of the theorem is trivial: if there is a p-structure π admissible with K', then there is at least a preferred model of p in π, in which q does not hold and, since π is also admissible with K, p cannot p-entail q in K. The 'if' part is slightly more involved unless K itself is p-inconsistent, in which case the proof is trivial. So let us assume that K is p-consistent and that p does not p-entail q in K. Then, there must be a p-structure $\pi = \langle \mathcal{I}, < \rangle$ in which there is a preferred

model M of p where q does not hold. In order to show that K' is p-consistent in that case, we construct a p-structure $\pi' = \langle \mathcal{I}, <' \rangle$ admissible with K' as follows. π' is defined by retaining the set of interpretations in π, and by *extending* the order '$<$' in such a way that $M_1 <' M_2$ if $M_1 < M_2$ for any interpretations M_1 and M_2 in $\mathcal{I}$, and $M <' M_3$, for any preferred model M_3 of p in π different than M. It is simple to check that π' is a p-structure admissible with K. Furthermore, M is the unique preferred model of p in π', and thus $\neg q$ is true in the non-empty set of preferred models of p in π'. It follows then that π' is admissible with K', and therefore, that K' is p-consistent. ∎

Lemma 3.2 *A default theory $T = \langle K, p \rangle$ with a background $K = \langle L, D \rangle$ l-entails a sentence q if only if the background $K' = \langle L, D \cup \{p \rightarrow \neg q\} \rangle$ is l-inconsistent.*

Proof Similar to the proof of lemma 3.1. ∎

Lemma 3.3 *A background K is p-consistent if and only if K is l-consistent.*

Proof We show first that given a l-structure $\lambda = \langle \mathcal{W}, \kappa \rangle$ admissible with a background context K, it is possible to construct a p-structure $\pi = \langle \mathcal{I}, < \rangle$ also admissible with K. For that purpose, we define $\mathcal{I}$ to be any minimal set of interpretations whose associated set of worlds is $\mathcal{W}$, and define the order '$<$' on interpretations in such a way that $M < M'$ holds for two interpretations M and M' in $\mathcal{I}$, iff $\kappa(w(M)) < \kappa(w(M'))$, where $w{:}\mathcal{I} \mapsto \mathcal{W}$, is a function that maps an interpretation into its corresponding world. We need to show that the structure π is admissible with $K = \langle L, D \rangle$ if λ is. First of all, note that due the fact that worlds in $\mathcal{W}$ have only non-negative ranks, the induced preferential structure π must be well-founded. Likewise, the mapping from worlds to interpretations preserves satisfiability; as the interpretations in $\mathcal{I}$ all satisfy L, and there is at least one which satisfies p, for every default $p \rightarrow q$ in D. Furthermore, if q is false in some preferred model M of p in $\mathcal{I}$, it means that there is a world W in which both p and $\neg q$ hold, and no world W' where p and q hold such that $\kappa(W') < \kappa(W)$. This, however, would contradict the assumption that the l-structure $\lambda = \langle \mathcal{W}, \kappa \rangle$ is admissible with K. Therefore, if λ is admissible, so is π.

To show that given a p-structure $\pi = \langle \mathcal{I}, < \rangle$ admissible with K it is possible to build a l-structure $\lambda = \langle \mathcal{W}, \kappa \rangle$ admissible with K, we use a construction suggested by Lehmann [1989]. First, we define $\mathcal{W}$ to be the set of worlds that

correspond to the interpretations in $\mathcal{I}$, and then for every interpretation M in $\mathcal{I}$ we let $height(M)$ stand for the length of the longest ascending chain of interpretations in $\mathcal{I}$, $M_0, M_1, \ldots, M_n$, such that $M_i < M_{i+1}$ for every i, $0 \leq i < n$, and where $M = M_n$. The rank of a world W in $\mathcal{W}$ is then defined as $\kappa(W) = \min_{M \in \mathcal{I}_W} height(M)$, where $\mathcal{I}_W$ stands for the set of interpretations M in $\mathcal{I}$ such that $w(M) = W$. Since, again, the mapping form worlds to interpretations preserves satisfiability, to show the admissibility of λ we need only to show that for every default $p \rightarrow q$ in K, q is true in all the preferred worlds of p in λ. Let us assume otherwise, that there is a world W in $\mathcal{W}$ that satisfies both p and $\neg q$, and no world W' in $\mathcal{W}$ that satisfies both p and q such that $\kappa(W') < \kappa(W)$. Furthermore, let $\kappa(W) = height(M)$, for some interpretation M in $\mathcal{I}$. Then, from the admissibility of the preferential model structure π, it must be the case that there is an interpretation M' that satisfies both p and q and for which $M' < M$. This in turn implies that $height(M') < height(M)$ and, therefore, that $W' = w(M')$ satisfies both p and q, and by construction $\kappa(W') < \kappa(W)$ also holds. This, however, contradicts the assumption that W is a preferred world of p in λ. Thus, λ is a layered structure admissible with K and therefore, K is l-consistent. ∎

Lemma 3.4 *A background K is ϵ-consistent if and only if K is l-consistent.*

Proof We show that given a l-consistent background $K = \langle L, D \rangle$ it is possible to construct a probability distribution admissible with K within any positive range, and vice versa, that it is possible to construct an admissible layered world structure given a ϵ-consistent background K.[1] Assume first that K is l-consistent and that $\langle \mathcal{W}, \kappa \rangle$ represents a layered structure admissible with K. As only a finite number worlds satisfy L, we assume that $\mathcal{W}$ contains n worlds and that the ranking κ divides the set $\mathcal{W}$ into $m+1$ non-empty layers $\mathcal{W}_0, \mathcal{W}_1, \ldots, \mathcal{W}_m$ of increasing rank, each with a number n_i of worlds. Clearly, since L has to be logically consistent, there is a probability distribution admissible with K within any range ϵ greater or equal than unity. We show below, that it is also possible to construct a probability distribution P over $\mathcal{W}$ which is admissible with K within *any* positive real ϵ smaller than unity.[2]

The probability distribution P is defined to assign to each world W in layer $\mathcal{W}_i$, $0 \leq i < m$, a probability $P(W) = \delta^i(1-\delta)/n_i$, $0 < \delta \leq \epsilon \cdot [1 + n(1-\epsilon)]^{-1} < 1$, and to each world W' in the last layer $\mathcal{W}_m$, a probability $P(W') = \delta^{m+1}$. Due to the equality between the expression $1 - \delta^{m+1}$ and its expansion $(1 + \delta + \delta^2 + \delta^3 +$

[1] See section 2.4 for the definition of admissible probability distributions in ϵ-semantics.

[2] Similar constructions appear in [Adams, 1966] and [Lehmann and Magidor, 1988].

$\ldots + \delta^m) \cdot (1 - \delta)$, it is easy to show that sum of P over all the worlds in $\mathcal{W}$ is 1. We need to show that that the probability distribution P is admissible with K within a range ϵ. First, note that since the structure $\langle \mathcal{W}, \kappa \rangle$ is admissible with K, every world W in $\mathcal{W}$ satisfies L, and for every default $p \rightarrow q$ in D there is a world where p holds. Thus, we are guaranteed that the probability $P(s)$ of any sentence s in L is one, and that the probability $P(p)$, for any default $p \rightarrow q$ in D, is greater than zero. We are thus left to show that the probability $P(q \mid p)$ for any default $p \rightarrow q$ in D is greater than $1 - \epsilon$. We know, however, due the admissibility of the structure $\langle \mathcal{W}, \kappa \rangle$, that there is a world W that satisfies both p and q which is better than any other world in which p is satisfied and q is not. In particular, if W belongs to $\mathcal{W}_m$, we are thus guaranteed $P(q \mid p) = 1$. Otherwise, W must belong to some layer $\mathcal{W}_i$, $0 \leq i < m$. In that case, we obtain that $P(q|p)$ must be equal or greater that $1 - \epsilon$ as follows:

$$
\begin{aligned}
P(q \mid p) &= \frac{P(p,q)}{P(p,q) + P(p,\neg q)} \\
&\geq \frac{\delta^i (1-\delta)}{\delta^i (1-\delta) + n\, \delta^{i+1}} \\
&\geq \frac{1-\delta}{1 + n\,\delta} \\
&\geq 1 - \epsilon \text{ , for } 0 < \delta \leq \frac{\epsilon}{1 + n\,(1-\epsilon)} < 1
\end{aligned}
$$

To prove the converse, we will construct an admissible layered-structure $\langle \mathcal{W}, \kappa \rangle$ given a ϵ-consistent background K. We will select the set of worlds $\mathcal{W}$ to be the finite set of worlds $W_1, \ldots, W_n$ which satisfy L. From the assumption that $\mathcal{L}$ is finite, we know that the set of possible worlds must be finite, and furthermore, that each world can be characterized in terms of the truth of a finite set of ground atoms. For a world W_i, we will refer by s_i to the sentence formed by conjoining the positive and negative ground literals true in W_i. Namely, s_i can be regarded as the 'world' sentence associated to W_i. The ranking κ on worlds will be determined, by defining an order on their corresponding sentences. First, we collect in a set S all the defaults of the form $s_i \vee s_j \rightarrow \neg s_i$ and $s_i \vee s_j \rightarrow \neg s_j$, for world sentences s_i and s_j with $i < j$. The number of defaults in S is thus $l = n\,(n-1)$. We then incrementally construct a new ϵ-consistent background context $K' = \langle L, D \cup S' \rangle$, $S' \subseteq S$, from the ϵ-consistent background context $K = \langle L, D \rangle$, as follows. Initially, we let $S^0 = S$, $D^0 = \emptyset$ and $K^0 = K$. Then until S^i is empty, for each $i = 1, \ldots, l$, we remove a default $p_i \rightarrow q_i$ from S^i and test whether $\neg q_i$ in ϵ-entailed by p_i in the

background context $K^{i-1} = \langle L, D \cup D^{i-1} \rangle$. If so, we set D^i to D^{i-1}; otherwise, we set D^i to $D^{i-1} \cup \{p_i \rightarrow q_i\}$. Note that the resulting background context $K' = K^l$ is ϵ-consistent; $K^0 = K$ is ϵ-consistent by assumption and, as a result of lemma 2.4, each iteration preserves the ϵ-consistency of K^i. We show next (1) that there is a *total* order over the sentences s_i, $i = 1, \ldots, l$, and (2) that the ranking κ on $\mathcal{W}$ determined by this order renders a layered world structure admissible with K.

Let s_i and s_j, $i < j$, stand for a pair of world sentences. We show first that one and only one of the defaults $s_i \vee s_j \rightarrow \neg s_i$ or $s_i \vee s_j \rightarrow \neg s_j$ belongs to K'. Assume that $s_i \vee s_j \rightarrow \neg s_i$ belongs to K'. Then the disjunction $s_i \vee s_j$ ϵ-entails the sentence s_j and, by consistency arguments, the default $s_i \vee s_j \rightarrow \neg s_j$ cannot belong to K'. Otherwise, if $s_i \vee s_j \rightarrow \neg s_i$ does not belong to K', it must be the case that it is not consistent with some K^i, $0 \leq i \leq l$, and therefore, not consistent with K'. That means, by lemma 2.4, that the disjunction $s_i \vee s_j$ ϵ-entails the sentence s_i in K', and therefore, that it also ϵ-entails the sentence $\neg s_j$, since s_i and s_j are logically inconsistent. Thus, the default $s_i \vee s_j \rightarrow \neg s_j$ is ϵ-consistent with K' and, therefore, it must belong to K'. Furthermore, from the soundness of core and **or-transitivity** (theorem 2.1), it must also be the case that if K' includes the defaults $s_i \vee s_j \rightarrow \neg s_j$ and $s_j \vee s_k \rightarrow \neg s_k$, then K' must also include the default $s_i \vee s_k \rightarrow \neg s_k$.

Thus, K' determines a total order '$<$' on the sentences s_i, where $s_i < s_j$ iff either $s_i \vee s_j \rightarrow \neg s_j$ or $s_j \vee s_i \rightarrow \neg s_j$ belong to K'. We can thus define the ranking κ of a world W_i as the length of the maximal chain $s_{i_1} < s_{i_2} < \ldots < s_{i_n}$, where $s_{i_n} = s_n$. We show now that the resulting l-structure $\lambda = \langle \mathcal{W}, \kappa \rangle$ is admissible with K. Note that by definition, $\mathcal{W}$ stands for the set of worlds that satisfy L. Furthermore, since $K = \langle L, D \rangle$ is ϵ-consistent, $\mathcal{W}$ must include worlds satisfying p for every default $p \rightarrow q$ in D. We are thus left to show that for every such default, q is true in all the preferred worlds of p in the l-structure λ. Assume otherwise, that there is a world W_i in which p holds and q does not, and no world W_j with smaller rank than W_i where both p and q hold. Since by construction κ orders the worlds in $\mathcal{W}$ along a chain, without loss of generality we can assume that W_i is the single minimal ranked world in $\mathcal{W}$ where p holds. However, if s_i is the sentence which corresponds to W_i, this requires K' to contain the defaults $s_j \vee s_i \rightarrow \neg s_j$ and $s_i \vee s_k \rightarrow \neg s_k$, for every sentence s_j and s_k consistent with p, for which $j < i$ or $i < k$. Furthermore, since p logically entails the disjunction of all such sentences s_k, s_j, and s_i, this implies, as a result of **or-monotonicity** (theorem 2.1), deductive closure, and the soundness of the core, that s_i is ϵ-entailed by p in K'. This, however would imply that $\neg q$ is ϵ-entailed by p in K' as well, in contradiction with the ϵ-consistency of K'. Thus, there is no such a world W_i and the l-structure λ is

admissible with K. ∎

Theorem 3.3 *Let $K = \langle L, D \rangle$, and $K' = \langle L, D + \{p \rightarrow \neg q\} \rangle$ be two background contexts, and let $T = \langle K, \{p\} \rangle$ be a default theory. Then, for finite propositional languages, the following statements are equivalent:*

(1) T ϵ-entails q
(2) K' is ϵ-inconsistent
(3) K' is l-inconsistent
(4) T l-entails q
(5) K' is p-inconsistent
(6) T p-entails q

Proof (1) and (2) are equivalent as a result of lemma 2.4; the same about (3) and (4) (lemma 3.2), and about (5) and (6) (lemma 3.1). Likewise, the equivalence between (2) and (3) follows from lemma 3.4, while the equivalence between (3) and (5) from lemma 3.3. ∎

Theorem 3.4 *A background context $K = \langle L, D \rangle$ is l-consistent if and only if there is a default ranking admissible with K.*

Proof We prove the 'only if' part first. Assume that K is l-consistent, and that $\lambda = \langle \mathcal{W}, \kappa \rangle$ in a layered world structure admissible with K. We show how to construct an admissible default ranking σ over $K = \langle L, D \rangle$, by choosing $\sigma(p \rightarrow q) = \min_{W \in \mathcal{W}_p} \kappa(W)$, where $\mathcal{W}_p$ stands for the non-empty set of worlds in $\mathcal{W}$ that satisfy p. We need to show that σ is a default ranking admissible with K. Assume it is not; i.e. there is a default $p \rightarrow q$ in D in conflict with a subset D' of D, such that $\sigma(p \rightarrow q) \leq \min_{\delta' \in D'} \sigma(\delta')$. Let us select W as a minimal ranked world in $\mathcal{W}$ that verifies the default $p \rightarrow q$. By the admissibility of λ, we know there is at least one such world, which we also know must satisfy L. From the conflict between $p \rightarrow q$ and D', one default in D', say $p' \rightarrow q'$, must then be falsified by W. Again, by the admissibility of λ, W must thus be preceded by a world W' in which the default $p' \rightarrow q'$ is verified. This, however, implies $\sigma(p' \rightarrow q') < \sigma(p \rightarrow q)$ in contradiction with the assumption above.

We prove now the 'if' part. We show that given an admissible default ranking over $K = \langle L, D \rangle$ it is possible to construct a layered structure $\langle \mathcal{W}, \kappa \rangle$ which is

admissible with $K = \langle L, D \rangle$. For that purpose, we select $\mathcal{W}$ as the set of all worlds consistent with L, and κ in such a way that $\kappa(W) = \max_{d \in D_W} \sigma(d)$, where D_W stands for the set of defaults falsified in world W. We need to show that for any world W in $\mathcal{W}$ which falsifies a default $p \rightarrow q$ in D, there is a world W', $\kappa(W') < \kappa(W)$, which verifies it. First, note that if W falsifies the default $p \rightarrow q$ then, by the definition of κ, $\kappa(W)$ must be equal or higher than $\sigma(p \rightarrow q)$. Moreover, we know from the fact that σ is an admissible default ranking, that every subset D_i of D in conflict with $p \rightarrow q$ contains a default $p_i \rightarrow q_i$ such that $\sigma(p_i \rightarrow q_i) < \sigma(p \rightarrow q)$. Let D' be the set of all such defaults. It follows then that there is world W' in $\mathcal{W}$ that satisfies L, verifies the default $p \rightarrow q$ and only falsifies defaults in D'. This implies that the ranking of W' is such that $\kappa(W') < \sigma(p \rightarrow q)$ and, therefore, that $\kappa(W') < \kappa(W)$. Thus, the layered world structure $\langle \mathcal{W}, \kappa \rangle$ is admissible with K, and K is l-consistent. ∎

Lemma 3.5 *A background context is l-consistent if and only if it does not contain a clash.*

Proof We prove the 'only if' part first. Let us assume that there is an admissible ranking σ over $K = \langle L, D \rangle$, and that a set D', $D' \subseteq D$ constitutes a clash in K. Let us further choose from D' a default $p \rightarrow q$ with a minimum rank, and let i represent its rank. There must be one such minimum ranked default at least, by the nature of default rankings. However, by the definition of *admissible* default ranking, since $p \rightarrow q$ is in conflict with D', D' must contain a default with a rank smaller than i, contradicting thus the minimality of $p \rightarrow q$. Thus, K cannot contain a clash.

Now, let us assume that K does not contain a clash. We show that it is possible construct an admissible default ranking by decomposing D into layers $D_0, D_1, \ldots, D_i, \ldots$ and by setting $\sigma(p \rightarrow q) = i$ iff $p \rightarrow q \in D_i$. Let $D^0 = D$. Since, in particular, D is not a clash in itself, there is a set D_0 of defaults in D^0 which are not in conflict with D^0. For $i = 1, 2, \ldots$, let D^i be set to $D^{i-1} - D_{i-1}$. Since $D^i \subset D$, D^i cannot be a clash and, therefore, there must be a non-empty set D_i of defaults in D^i which are not in conflict with D^i. Following this procedure, we obtain a layering $D_0, D_1, \ldots$ of defaults, such that (1) $D = \cup_i D_i$, and (2) every default in D_i is not conflict with defaults in $D^i = D_i \cup D_{i+1} \cup \cdots \cup D_n$. Thus, since D contains a finite number of default schemas whose instances all belong to the same layer, the default ranking $\sigma(p \rightarrow q) = i$ iff $p \rightarrow q \in D_i$ assigns a rank to every default in D, and therefore, σ is default ranking admissible with K. ∎

Lemma 3.6 *p entails q in $K = \langle L, D\rangle$ if and only if the background $K' = \langle L, D + \{p \rightarrow \neg q\}\rangle$ contains a clash.*

Proof Follows from the fact that p entails q in K iff K' is inconsistent, which in light of lemma 3.5 and the equivalences between l-consistency, p-consistency, and ϵ-consistency, amounts to the presence of a clash in K'. ∎

Theorem 3.5 *For a background context $K = \langle L, D\rangle$ with n defaults, there is an $\mathcal{O}(C(n) \times n^2)$ procedure for testing whether a sentence q is entailed by a sentence p in K, where $C(n)$ is the complexity associated with classical entailment in the language fragment that comprises the sentences of L and the material counterparts of the defaults in D.*

Proof In order to test whether q is entailed by p, it is sufficient to test the consistency of the background K' as in lemma 3.6. We can test the consistency of K' by simply following the construction given in the second part of the proof of lemma 3.5. That is, we start with $D^0 = D$ and for each D^i we identify a set D_i of defaults which are not in conflict with D^i, and set $D^{i+1} = D^i - D_i$. We stop this iteration only when either one of the sets D^i or D_i is empty. If D^i is empty at the end, it means that we have not found a clash in D, and therefore, that K' is consistent and that p does not entail q in K. Otherwise, D^i *is* a clash and, therefore, K' is inconsistent and q is entailed by p. Furthermore, there are at most $n + 1$ iterations, each involving at most $n + 1$ satisfiability tests. ∎

Lemma 3.7 *Let $p \rightarrow q$ be a default in D, let D' be a subset of D, and let $C(D')$ be the quasi-conjunction of D'. Then, $p \rightarrow q$ clashes with D' in a background context $K = \langle L, D\rangle$, if and only if $p \rightarrow q$ clashes with $C(D')$ in the background context $K' = \langle L, D''\rangle$, with $D'' = \{C(D'), p \rightarrow q\}$.*

Proof Let $p_0 = p$, $q_0 = q$, and let D' be the collection of defaults $p_i \rightarrow q_i$, $i = 1, \ldots, n$. First note that, if the set composed by $p_0 \rightarrow q_0$ and D' is a clash in K, we must have

$$p_i \vdash_K \neg(p_0 \Rightarrow q_0) \vee \neg(p_1 \Rightarrow q_1) \vee \cdots \vee \neg(p_n \Rightarrow q_n)$$

for each $i = 0, 1, \cdots, n$. Therefore, by invoking the logical equivalence between the formulas:

$$\neg(p_1 \Rightarrow q_1) \vee \cdots \vee \neg(p_n \Rightarrow q_n)$$

and

$$\neg(p_1 \vee p_2 \vee \cdots \vee p_n \Rightarrow (p_1 \Rightarrow q_1) \wedge (p_2 \Rightarrow q_2) \wedge \cdots \wedge (p_n \Rightarrow q_n))$$

we get

$$p_0 \mathrel{\mid\!\sim_K} \neg(p_0 \Rightarrow q_0) \vee \neg(p_1 \vee p_2 \vee \cdots \vee p_n \Rightarrow (p_1 \Rightarrow q_1) \wedge \cdots \wedge (p_n \Rightarrow q_n)),$$

and

$$p_1 \vee \cdots \vee p_n \mathrel{\mid\!\sim_K} \neg(p_0 \Rightarrow q_0)\neg(p_1 \vee p_2 \vee \cdots \vee p_n \Rightarrow (p_1 \Rightarrow q_1) \wedge \cdots \wedge (p_n \Rightarrow q_n)),$$

where the last derivation involves the logical equivalence between $x \vee y \mathrel{\mid\!\sim_K} z$ and $x \mathrel{\mid\!\sim_K} z$ and $y \mathrel{\mid\!\sim_K} z$. The last two expressions above reveal a clash between the default $p_0 \rightarrow q_0$ and quasi-conjunction $C(D')$. The proof for the 'if' part of the lemma involves the reverse steps. ∎

Lemma 3.8 *Let $K = \langle L, D\rangle$ be a background context, and D' be a non-empty subset of D. Then, if $r \rightarrow s$ stands for the quasi-conjunction $C(D')$ of D', $r \mathrel{\mid\!\sim^{\circ}_K} s$.*

Proof Let D' be a collection of defaults $p_i \rightarrow q_i$, $i = 1, \ldots, n$. Then by rule 1 of the core we can obtain $p_i \mathrel{\mid\!\sim^{\circ}_K} q_i$. Furthermore, if let $p_{1,n}$ stand for the disjunction $p_1 \vee \cdots \vee p_n$, we can get $p_i, p_{1,n} \mathrel{\mid\!\sim^{\circ}_K} q_i$ by augmentation, and $p_{1,n} \mathrel{\mid\!\sim^{\circ}_K} p_i \Rightarrow q_i$, for any $i = 1, \ldots, n$, by weak reduction. Finally, the target result $p_{1,n} \mathrel{\mid\!\sim^{\circ}_K} (p_1 \Rightarrow q_1) \wedge \cdots \wedge (p_n \Rightarrow q_n)$ follows by deductive closure. ∎

Lemma 3.9 *Let $K = \langle L, D\rangle$, and $K' = \langle L, D'\rangle$ be two background contexts sharing the same set L of sentences. If $p \rightarrow \neg q$ clashes with $r \rightarrow s$ in K' and $r \mathrel{\mid\!\sim^{\circ}_K} s$, then $p \mathrel{\mid\!\sim^{\circ}_K} q$.*

Proof From the definition of default clashes, it follows by **deduction** that $r, s \mid\!\sim^{\circ}_{K} p \wedge q$ and $p, \neg q \mid\!\sim_{K} r \wedge \neg s$. Furthermore, from $r \mid\!\sim^{\circ}_{K} s$ and the first expression, we can obtain $r \mid\!\sim^{\circ}_{K} p$, $r, p \mid\!\sim^{\circ}_{K} q$, and therefore, by **weak reduction**, $p \mid\!\sim^{\circ}_{K} \neg r \vee q$. Likewise, from the second expression we can obtain $p, \neg q \mid\!\sim^{\circ}_{K} r$, and then by **weak reduction**, $p \mid\!\sim^{\circ}_{K} r \vee q$. Combining the two results by means of **deductive closure**, $p \mid\!\sim^{\circ}_{K} q$ thus follows. ∎

Lemma 4.1 *The quadruple* $\langle \mathcal{I}_{\mathcal{L}}, <, \Delta_{\mathcal{L}}, \prec \rangle$ *is a prioritized preferential structure only if the pair* $\langle \mathcal{I}_{\mathcal{L}}, < \rangle$ *is a preferential structure.*

Proof We know that for two interpretations M and M', the relation $M < M'$ holds iff $\Delta[M] \neq \Delta[M']$, and for every δ in $\Delta[M] - \Delta[M']$, there exists a δ' in $\Delta[M'] - \Delta[M]$, such that $\delta \prec \delta'$, where '$\prec$' is an irreflexive and transitive relation which does not contain infinite chains. First, note that the relation '$<$' is clearly irreflexive. We need to show that '$<$' is also transitive. Let M_1, M_2, and M_3 be three interpretations such that $M_1 < M_2$ and $M_2 < M_3$, and let $\Delta_1 = \Delta[M_1]$, $\Delta_2 = \Delta[M_2]$, and $\Delta_3 = \Delta[M_3]$. We will use the notation $\overline{\Delta}$ to denote the complement of a set Δ , i.e. $\overline{\Delta} = \Delta_{\mathcal{L}} - \Delta$. Moreover, we will find convenient to denote the intersection of sets Δ of assumptions with indices i_1, i_2, ..., i_n, by simply writing $\Delta_{i_1,i_2,\ldots,i_n}$. Furthermore, when one of the indices i is preceded by a minus sign, the associated assumption set Δ_i is supposed to be replaced by its complement $\overline{\Delta_i}$. Thus, for instance $\Delta_{1,-2,3}$ stands for the intersection of the sets Δ_1, Δ_3 and the complement $\overline{\Delta_2}$ of Δ_2. Similarly, $\Delta_{-1,2}$ stands for the intersection of $\overline{\Delta_1}$ and Δ_2.

We need to show that for every assumption δ in $\Delta_{1,-3}$, there is an assumption δ' in $\Delta_{-1,3}$ such that $\delta \prec \delta'$.[3] Note that it is sufficient to prove this for every *maximal* element δ in $\Delta_{1,-3}$. Since '$\prec$' does not contain infinite chains, it is clear that for every non-maximal element δ'' in $\Delta_{1,-3}$ there is a maximal element δ such that $\delta'' \prec \delta$. So, if $\delta \prec \delta'$ holds, by transitivity $\delta'' \prec \delta'$ will hold as well. Hence, let δ_1 be an arbitrary maximal element in $\Delta_{1,-3}$. We need to consider two main cases:

1. if δ_1 belongs to $\Delta_{1,-2,-3}$, then δ_1 must also belong to $\Delta_{1,-2}$. Thus, since $M_1 < M_2$, there must be a δ_2 in $\Delta_{-1,2}$ such that $\delta_1 \prec \delta_2$. Furthermore, let δ_2 the maximal such element. If $\delta_2 \in \Delta_{-1,2,3}$ we are done. Otherwise, $\delta_2 \in \Delta_{-1,2,-3}$ and then, since $M_2 < M_3$, there must be a $\delta_3 \in \Delta_{-2,3}$ such

[3] An equivalent proof can be found in [Przymusinski, 1987].

that $\delta_2 \prec \delta_3$. If $\delta_3 \in \Delta_{1,-2,3}$, then from $M_1 < M_2$, there should be a δ_2' in $\Delta_{-1,2}$ such that $\delta_3 < \delta_2'$, and therefore, $\delta_2 \prec \delta_2'$, in contradiction with the maximality of δ_2. Thus, $\delta_3 \in \Delta_{-1,3}$, and $\delta_1 \prec \delta_3$, by the transitivity of '$\prec$'.

2. if δ_1 belongs to $\Delta_{1,2,-3}$, then, since $M_2 < M_3$, there must be a δ_3 in $\Delta_{-2,3}$ such that $\delta_1 \prec \delta_3$. Moreover, if $\delta_3 \in \Delta_{-1,-2,3}$ we are done. Otherwise, $\delta_3 \in \Delta_{1,-2,3}$, and therefore, as a result of $M_1 < M_2$, there must be a δ_2 in $\Delta_{-1,2}$ such that $\delta_3 \prec \delta_2$. Let δ_2 be a maximal such element. if δ_2 belongs to Δ_3 we are done. Otherwise, $\delta_2 \in \Delta_{-1,2,-3}$, and therefore, there must be a δ_3' in $\Delta_{-2,3}$ such that $\delta_2 \prec \delta_3'$. Furthermore, δ_3' cannot belong to Δ_1; otherwise, there should be another element δ_2' in $\Delta_{-1,2}$, such that $\delta_3' \prec \delta_2'$, contradicting the maximality of δ_2. So, $\delta_3' \in \Delta_{-1,3}$ and $\delta_1 \prec \delta_3'$ by transitivity of '$\prec$.' ∎

Lemma 4.2 *For two models M and M' of a theory T, if $\Delta[M] \subset \Delta[M']$, then M is preferred to M' ($M < M'$) in every prioritized preferential structure.*

Proof If $\Delta[M] \subset \Delta[M']$, then $\Delta[M] - \Delta[M'] = \emptyset$, and the relation $M < M'$ holds trivially in every prioritized preferential structure. ∎

Lemma 4.3 *If M is a preferred model of a theory T in a given induced preferential structure, then M is* minimal *in $\Delta_{\mathcal{L}}$, i.e. there is no model M' of T such that $\Delta[M'] \subset \Delta[M]$.*

Proof Straightforward from lemma 4.2. ∎

Lemma 4.4 *Bound default theories are well-founded.*

Proof In order to show that a bound theory T is well-founded, we appeal to the notion of *hitting sets* in [Reiter, 1987b]. For a collection C of sets $\Delta_1, \ldots, \Delta_n$, a set Δ is *a hitting set* for C iff Δ includes an assumption δ from every set Δ_i in C. If we let C stand for the minimal conflict sets in T, then a necessary condition for an interpretation M to be a model of T, is for the gap $\Delta[M]$ of M to include a hitting set for C. Moreover, a model of T will be minimal iff $\Delta = \Delta[M]$ is a *minimal* hitting set for C. Also note that a bound theory T gives rise to a finite set of minimal classes, and since the gap of every non-minimal class must contain

a minimal hitting set for C, every class is either minimal or has a gap larger than some minimal class.

Now, if T is logically inconsistent, the lemma follows trivially. So let us assume that T is logically consistent and that it gives rise to a finite number n of minimal classes. If a minimal model M_1 of T is not a preferred model of T in some prioritized structure π, there must be another minimal model M_2 of T which is preferred to M_1. If M_2 in turn is not a preferred model of T in π, then there has to be another minimal model M_3 of T preferred to M_2 and so on. However, one such model M_i for $i \leq n$ *must* be a preferred model of T in π because in the presence of n minimal classes there cannot be a chain containing more than n minimal models. So, if M is a minimal model of T, then either M is a preferred model of T in π, or there must be a preferred model M' of T such that $M' < M$. Similarly, for a non-minimal model M'' of T it follows from the remarks above that there must be a minimal model M of T such that $\Delta[M] \subseteq \Delta[M'']$, and therefore $M < M'$. Thus from the result above, either M is a preferred model of T, or there is a preferred model M' of T such that $M' < M < M''$. ∎

Theorem 4.1 *For a theory $T = \langle K, E \rangle$ over a finite propositional language, T preferentially entails p only if T conditionally entails p.*

Proof Note that if $\mathcal{L}$ is finite propositional language the p-structure $\pi = \langle \mathcal{I}_\mathcal{L}, < \rangle$ embedded in any prioritized structure $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$ must be well-founded. We further show that π is also admissible with K and hence, that if T does not conditionally entail p, T does not preferentially entail p either. For that we need to prove that for every default $p \rightarrow \delta$ in D, δ holds in the preferred models of the theory $T' = \langle K, \{p\} \rangle$ in π. Let M' be a model of T' in which δ does not hold. Thus, clearly, $\delta \in \Delta[M']$. We construct a model M preferred to M' in which δ holds. Since the preference order '$<$' is well-founded, this is sufficient to prove that δ holds in all preferred models of T'. Let C stand for the collections of all minimal conflict sets in T, and let C' stand for the collection of all minimal conflict sets Δ in T such that $\Delta \cap \Delta[M'] = \{\delta\}$. Since the priority ordering '$\prec$' is admissible, any such set Δ must contain an assumption δ' such that $\delta' \prec \delta$. Let Δ' be the collection of all such assumptions δ', and let us select M as an interpretation which satisfies T, with a gap $\Delta[M] = \Delta[M'] + \Delta' - \{\delta\}$. From the results in the proof of lemma 4.4 above, there must be one such interpretation as $\Delta[M]$ is a hitting set for C. Indeed, any set in C not 'hit' by assumptions in $\Delta[M'] - \{\delta\}$; will certainly be 'hit' by assumptions in Δ'. Thus, the relation $M < M'$ must hold, as

$\Delta[M] - \Delta[M'] = \Delta'$, $\Delta[M'] = \Delta[M] = \{\delta\}$, and for every δ' in Δ, the relation $\delta' \prec \delta$ holds. ∎

Theorem 4.2 *For a pure background context K over a finite propositional language, K is p-consistent only if K is cd-consistent.*

Proof The proof below relies on the correspondence between the p-consistency of K and the presence of default rankings admissible with K (theorems 3.4 and 3.3, chapter 3), as in the particularity of *pure* theories, which permit a natural translation of admissible default rankings σ into admissible priority orderings '$\prec$'. Indeed, if $p_i \rightarrow \delta_i$, $i = 1, 2, \ldots, n$ are the defaults in a pure background K, the condition:

$$p_i \mid\!\!\sim_K \neg(p_1 \Rightarrow \delta_1) \vee \neg(p_2 \Rightarrow \delta_2) \vee \cdots \vee \neg(p_n \Rightarrow \delta_n)$$

for some i, $1 \leq i \leq n$, is equivalent to the condition:

$$p_i \mid\!\!\sim_K \neg\delta_1 \vee \neg\delta_2 \vee \cdots \vee \neg\delta_n$$

Therefore, if σ is a default ranking admissible with K, the priority ordering '$\prec$' defined as $\delta_i \prec \delta_j$ iff $\sigma(p_i \rightarrow \delta_i) < \sigma(p_j \rightarrow \delta_j)$, will also be admissible with K. ∎

Lemmas 4.6, 4.7, and 4.8, are special cases of the following theorem:

Theorem 4.3 *An assumption δ is conditionally entailed in a context T if and only if δ belongs to a stable cover in T.*

Proof (only if part) We assume that $T = \langle K, E\rangle$ is a bound theory, and therefore, that T is well-founded and there are only a finite number of preferred classes of T. Let Δ_1, Δ_2, ..., Δ_n, be the *maximal* sets of assumptions validated in each of the preferred classes of T. Clearly, δ belongs to every such set. We show first that the collection C of sets (arguments) $\Delta_1, \ldots, \Delta_n$ constitutes a stable cover. More precisely, we show that if C is not a stable cover, there must be a preferred model of T in which none of the assumption sets in C holds. Let us thus assume that there is a set of assumptions Δ' which is in conflict with each of the sets $\Delta_1, \ldots, \Delta_n$, such that no set Δ_i, $1 \leq i \leq n$, is strongly protected from Δ'. That means that there is a priority ordering '$\prec$' admissible with K, such that for every i, $i = 1, \ldots, n$, a subset Δ_i'' of Δ' in conflict with Δ_i can be found, such that $\Delta_i'' \not\prec \Delta_i^j$ holds for every set Δ_i^j in Δ_i in conflict with Δ_i''. Furthermore, let δ_i^j be the assumption in Δ_i^j

for which the relation $\Delta_i'' \prec \delta_i^j$ fails to hold. With these assumptions we construct a model M of T which validates all the assumptions in Δ' and which among the sets in the cover C, only falsifies the assumptions δ_i^j. There is one such model, as every set Δ_i in C contains one assumption δ_i^j for every subset of Δ_i in conflict with Δ'.

Now, M cannot be a preferred model of T, since M violates one assumption of every set $\Delta_1, \ldots, \Delta_n$. Thus, since the theory is well-founded, there must be a preferred model M' of T such that $M' < M$. Let us assume that M' is a model which satisfies every assumption in one of the sets Δ_i, $1 \leq i \leq n$. Since Δ_i is in conflict with the subset Δ_i'' of Δ', M' must falsify one of the assumptions in Δ_i''. That is, some assumption δ_i'' in Δ_i'' must belong to $\Delta[M'] - \Delta[M]$. The preference of M' over M, then requires the set $\Delta[M] - \Delta[M']$ to contain an assumption δ such that $\delta_i'' \prec \delta$. However, $\Delta[M] - \Delta[M']$ only contains the assumptions δ_i^j, $i = 1, \ldots, n_i$, selected in a way such that the relation $\Delta_i'' \prec \delta_i^j$ does not hold. Thus, there cannot be preferred model M', $M' < M$, which satisfies all the assumptions in Δ_i, $1 \leq i \leq n$, and since M cannot be a preferred model of T, C must be a stable cover in T.

Proof (if part) We show now that if δ belongs to a stable cover $\Delta_1, \Delta_2, \ldots, \Delta_n$, in a bound theory $T = \langle K, E \rangle$, then δ is conditionally entailed by T. Since T is a well-founded theory it is sufficient to show that in any structure $\langle \mathcal{I}_\mathcal{L}, <, \Delta_\mathcal{L}, \prec \rangle$admissible with K, for any model M which violates assumptions from every set Δ_i, $i = 1, \ldots, n$, there is another model M' which satisfies one of the sets Δ_i, $1 \leq i \leq n$, such that $M' < M$. Let Δ' be the maximal set of assumptions satisfied by M. Clearly, if there is a set Δ_i, $1 \leq i \leq n$, not in conflict with Δ' in T, then there must be a model M' of T which satisfies both Δ' and Δ_i, and therefore, which is preferred to M. Let us assume otherwise, that Δ is in conflict with every set in the cover. Then, by the definition of stable cover, one of the sets Δ_i must be strongly protected from Δ'. That is, for every subset Δ_j' of Δ' in conflict with Δ_i, there is a subset Δ_i^j of Δ_i in conflict with Δ', such that $\Delta_j' \prec \Delta_i^j$. That means that every set Δ_j' in Δ' in conflict with Δ_i, contains an assumption δ_j', such that $\delta_j' \prec \delta_i^j$, for an assumption δ_i^j in in $\Delta[M]$. Let Δ'' stand for the collection of those assumptions δ_j' in Δ'. Then, it is possible to build a model M' of T that satisfies Δ_i such that $\Delta[M'] - \Delta[M] \subseteq \Delta''$, and thus, for which the relation $M' < M$ must hold. ∎

Theorem 4.4 (Main) *A proposition p is conditionally entailed in a context $T = \langle K, E \rangle$ if and only if p is supported by a stable cover in T.*

Proof Let $\Delta_1, \ldots, \Delta_n$ stand for the maximal assumptions sets legitimized by each of the preferred classes of T. From the proof of theorem 4.3 above, we know that such a collection of sets constitute a stable cover in T. Furthermore, if p is conditionally entailed by T, this means that T together with any of the sets Δ_i, $1 \leq i \leq n$, logically implies p. So, p is indeed supported by a stable cover. Let as assume now that p is supported by a stable cover $\Delta_1, \ldots, \Delta_n$ in T. Again, from the proof of theorem 4.3 above, we know that for every model M of T which satisfies no set Δ_i, there is a model M' of T which does satisfy one such set, and hence, since every such sets supports p, a model M' which satisfies p. Thus, since we are assuming T to represent a bound, and therefore, a well-founded theory, this amounts to say that p holds in all the preferred models of T. ∎

Theorem 4.5 *For two sets of assumptions Δ and Δ', the relation $\Delta' \prec \Delta$ holds in every priority ordering '$\prec$' admissible with a consistent background $K = \langle L, D \rangle$ if and only if Δ is included in a set that* dominates *the set of assumptions Δ' in K.*

Proof Let us recall, that we use the notation $\Delta' \prec \Delta$ to state that for every δ in Δ there exists a δ' in Δ' such that $\delta' \prec \delta$ holds. Moreover, the relation '$\prec$' among *sets* of assumptions remains irreflexive and transitive, and therefore, asymmetric. That is, for every priority ordering $\Delta \not\prec \Delta$, and if $\Delta_1 \prec \Delta_2$ and $\Delta_2 \prec \Delta_3$ holds, so does $\Delta_1 \prec \Delta_3$.

Let Δ stand for a collection of assumptions δ_i, $i = 1, \ldots, n$. We will use the notation $\Delta_{i,j}$, for $i \leq j$, to stand for the set $\{\delta_i, \delta_{i+1}, \ldots, \delta_j\}$. If $j > n$, the notation $\Delta_{i,j}$ is to be understood as $\Delta_{i,n}$, and if $i > n$, $\Delta_{i,j}$ denotes the empty set. We show that if Δ dominates a set Δ' then the relation $\Delta' \prec \Delta$ must hold for any priority ordering '$\prec$' admissible with K. We show this by induction; the base case $\Delta_{2,n} + \Delta' \prec \Delta_{1,1}$, first. Clearly, if Δ dominates Δ', the assumption δ_1 must d-dominate $\Delta_{2,n} + \Delta'$, and thus, $\Delta_{2,n} + \Delta' \prec \delta_1$ must hold. In particular, if $n = 1$, we are done. So let us assume that n is greater than one. Furthermore, let us assume as inductive hypothesis that $\Delta_{i+1,n} + \Delta' \prec \Delta_{1,i}$ holds for every i, $1 \leq i < n$. We need to show the same relation for $i = n$, for which $\Delta_{n+1,n} = \emptyset$ and $\Delta_{1,n} = \Delta$. By hypothesis, we have that $\{\delta_n\} + \Delta' \prec \Delta_{1,n-1}$, since $\Delta_{n,n} = \{\delta_n\}$. Let Δ_A stand for the set of assumptions in $\Delta_{1,n-1}$, such that $\{\delta_n\} \prec \Delta_A$ holds, and let Δ_B stands for $\Delta_{1,n-1} - \Delta_A$. Then, since the assumption δ_n d-dominates $\Delta + \Delta'$, there must an assumption δ' in $\Delta + \Delta'$, such that $\delta' \prec \delta_n$. Furthermore, δ' cannot belong to Δ_A, because $\{\delta_n\} \prec \Delta_A$, and the relation $\prec$ is asymmetric. So there are two cases to consider. If δ' belongs to the set Δ', then by transitivity we

would have $\Delta' \prec \Delta_A$, and therefore, $\Delta' \prec \Delta_{1,n}$. On the other hand, if $\delta' \in \Delta_B$ the relation $\Delta' \prec \delta_n$ must hold by transitivity on Δ_B, since the way Δ_B was selected guarantees $\Delta' \prec \Delta_B$ to hold. Furthermore, by transitivity on δ_n, $\Delta' \prec \Delta_A$ must hold as well, and therefore, $\Delta' \prec \Delta_{1,n}$ and $\Delta' \prec \Delta$ must hold as well.

The proof for the 'only if' part of the theorem is slightly more involved. We need to show that if the relation $\Delta' \prec \Delta$ holds for every admissible ordering with a (conditionally) consistent background context K, then Δ is part of a set that dominates Δ'. Let us first divide the assumptions in $\Delta_{\mathcal{L}}$ between those which participate in a set that dominates Δ', which we group in a set Δ_A, from those which do not participate in a set that dominates Δ'. Furthermore, let $\Delta_B = \Delta' - \Delta_A$, and $\Delta_C = \Delta_{\mathcal{L}} - \Delta_A - \Delta_B$. Note that Δ_B cannot be empty, otherwise Δ_A would dominate itself, precluing K from being consistent. Note also, that if two sets dominate Δ', so will their union. It follows then that Δ_A dominates Δ'. Our goal will be to show that Δ is included in Δ_A. For that we will show that there is a priority ordering '$\prec$' admissible with K, such that the relation $\Delta' \prec \delta$ holds only if $\delta \in \Delta_A$.

Let us say that a priority ordering '$\prec$' in a background context K is admissible within a *range* Δ and a *restriction* Δ' iff every set Δ'' d-dominated by an assumption δ in Δ contains an assumption δ' in Δ', such that $\delta' \prec \delta$ holds. The notions of *range* and *restriction* provide a finer measure of the admissibility of a priority ordering. In particular, an admissible priority ordering, must be admissible within a range $\Delta_{\mathcal{L}}$ and a restriction $\Delta_{\mathcal{L}}$. Furthermore, if a priority relation '$\prec$' is admissible within a range Δ_1 and a restriction Δ_2, for two sets Δ_1 and Δ_2 such that $\Delta_1 + \Delta_2 = \Delta_{\mathcal{L}}$, then there must be a priority relation '$\prec$' admissible within a range Δ_1 and restriction $\Delta_{\mathcal{L}}$, such that $\delta_2 \prec \delta_1$ holds only if $\delta_1 \in \Delta_1$ and $\delta_2 \in \Delta_2$. Indeed, if '$\prec'$' is a priority relation admissible within a range Δ_1 and a restriction Δ_2, the relation that results by deleting all pairs $\delta_1 \notin \Delta_1$ and $\delta_2 \notin \Delta_2$ for which $\delta_1 \prec' \delta_2$ holds, remains irreflexive, transitive, and admissible.

Now, let us assume that there is no priority ordering admissible within a range Δ_C and a restriction Δ_C, for Δ_C as above. By arguments similar to the ones about default clashes in section 3.5, it is possible to show then, that there must be a non-empty subset Δ'_C of Δ_C such that each assumption $\delta' \in \Delta'_C$ d-dominates the set $\Delta'_C + \overline{\Delta_C}$, where $\overline{\Delta_C}$ stands for the set of assumptions not in Δ_C; in this case, $\Delta_A + \Delta_B$. This, however, amounts to say that Δ'_C *dominates* the set $\Delta_A + \Delta_B$, which by virtue of the dominance of Δ_A over Δ' and the inclusion of Δ_B in Δ', implies that Δ'_C dominates Δ' as well, in contradiction with the maximality of Δ_A. Thus, there must be a priority ordering '$\prec_C$' admissible within a range Δ_C

and a restriction Δ_C, such that $\delta \prec_C \delta'$ holds only if δ and δ' both belong to Δ_C. Furthermore, since K is consistent, there must be a priority ordering '$\prec_A$' admissible within range Δ_A and restriction $\Delta_\mathcal{L}$, such that $\delta \prec_A \delta'$ holds only if δ' belongs to Δ_A. We can thus define a relation '$\prec$' such that $\delta \prec \delta'$ iff $[\delta \prec_A \delta']$ or $[\delta \prec_C \delta']$ or $[\delta \in \Delta_C$ and $\delta' \in \Delta_A + \Delta_B]$. It is simple to show that such a relation is a priority relation, and that it is admissible within a range $\Delta_A + \Delta_C$. Let us assume, on the other hand, that '$\prec$' is not admissible within a range Δ_B. That is, there is an assumption δ in Δ_B which d-dominates a set Δ'_B for which the relation $\Delta'_B \prec \delta$ fails to hold. Note that Δ'_B cannot contain elements from Δ_C; for, otherwise, the relation Δ'_B will certainly hold. Thus, $\Delta'_B \subseteq \Delta_A + \Delta_B$, so that δ d-dominates $\Delta_A + \Delta_B$. That means, however, that the set $\Delta_A + \{\delta\}$ dominates the set Δ', in contradiction with the assumption that Δ_A is the maximal such set. So, the ordering '$\prec$' must be admissible within the range Δ_B as well, and so '$\prec$' must also be a priority relation admissible with K. Since $\Delta' \prec \Delta$ holds by hypothesis, and $\Delta' \prec \delta$ holds only if $\delta \in \Delta_A$, it follows that Δ belongs to a set, Δ_A, which dominates Δ'. ∎

Theorem 4.6 *An assumption δ is conditionally entailed in a a context T, if for every argument Δ' against δ, there is a set Δ, $\delta \in \Delta$, that dominates Δ'.*

Proof The dominance Δ over Δ', implies $\Delta' \prec \delta$, for every assumption δ in Δ and every admissible priority ordering '$\prec$' (theorem 4.5). The assumption δ is thus protected from every conflicting set Δ' in T, and thus by lemma 4.8, δ is conditionally entailed by T. ∎

Theorem 4.7 *For finite propositional languages,* all *the rules of* **P** *are sound rules of conditional entailment.*

Proof We have shown in chapter 3 that rules 1–5 of **P** are sound with respect to preferential entailment (theorem 3.2), and in this chapter that preferential entailment is sound with respect to conditional entailment for finite propositional languages (theorem 4.1). The theorem thus follows from the soundness of the irrelevance rule established in theorem 4.6. ∎

In the proofs below, we use the symbols N, N' and N'' to stand for sets of *non-causal* atoms, and $C[N]$, for $N = \cup_i\{\alpha_i\}$, to stand for the set of *causal* atoms

$C[N] = \cup_i\{\mathrm{C}\alpha_i\}$. A set of atoms $C[N] + N'$ will thus denote a Herbrand interpretation which satisfies the causal atoms in $C[N]$ and the non-causal atoms in N'. We use the notation R_N, for a collection of rules R, to stand for the set of rules in R which contain no negative literal $\neg\alpha$ for no atom α in N. Likewise, R^+_N will denote the collection of rules that result from removing the negative literals from the bodies of the rules left in R_N. Finally, $\mathcal{C}[T; M]$ will denote the class of Herbrand models of T with an atomic gap M. Namely $\mathcal{C}[T; M]$ will stand for the collection of models $C[N] + N'$ of T such that $N' \subseteq M$.

Lemma 5.1 *M is a* stable *model of an arbitrary program P if and only if C_M is a perfectly coherent class of the causal theory $C_1[P]$.*

Proof We prove the lemma by showing the equivalence between the following conditions:

1. M is stable model of P
2. M is a minimal model of P^+_M
3. $C[M] + M$ is a minimal model of $C_1[P]^+_M$
4. the class $\mathcal{C}[C_1[P]^+_M; M]$ is perfectly coherent
5. the class $\mathcal{C}[C_1[P]; M]$ is perfectly coherent

The correspondence between (1) and (2) is the definition of stable models [Gelfond and Lifschitz, 1988]. For the correspondence between (2) and (3), note that a Herbrand interpretation N is a model of P^+_M iff $C[N] + N'$, for some set $N' \supseteq N$, is a model of $C_1[P]^+_M$.[4] Indeed, the theory $C_1[P]^+_M$ only contains rules of the form $\mathrm{C}\alpha_1 \wedge \cdots \wedge \mathrm{C}\alpha_n \Rightarrow \mathrm{C}\gamma$, for positive literals α_i and γ and, furthermore, for each such rule there is rule $\gamma \leftarrow \alpha_1, \ldots, \alpha_n$ in P^+_M, and vice versa. The equivalence between (2) and (3) thus follows: N is a *minimal* model of P^+_M iff $C[N] + N$ is a *minimal* model of $C_1[P]^+_M$. Furthermore, $C[N] + N$ is a minimal model of $C_1[P]^+_M$ iff the class $\mathcal{C}[C_1[P]^+_M; N]$ is perfectly coherent. Indeed, for any model $C[N'] + N''$ in $\mathcal{C}[C_1[P]^+_M; N]$, the relations $N'' \subseteq N$ and $N' \subseteq N''$ must hold. The first, due to gap of the class; the second due to the constraint on the causal operator C. Hence, if $C[N] + N$ is a *minimal* model, the relation $C[N] \subseteq C[N']$ must hold as well.

[4] We are assuming here that the only constraint on the causal operator C, is [C1]; namely, that every model of $\mathrm{C}\alpha$ is also a model of α.

Otherwise, we will get $C[N'] \subset C[N]$ and that the model $C[N]+N$ is not minimal. Likewise, if for every model $C[N']+N''$ in $\mathcal{C}[C_1[P]^+_M; N]$ the relation $C[N] \subseteq C[N']$ holds, then $C[N] + N$ must be a minimal model. Thus, we have the equivalence between (3) and (4). We are left then to show that the class $\mathcal{C}[C_1[P]^+_M; M]$ is perfectly coherent whenever the class $\mathcal{C}[C_1[P]; M]$ is. For that it is sufficient to show that both classes contain the same models. First, since perfectly coherent classes are minimal, models in either class $\mathcal{C}[C_1[P]^+_M; M]$ or $\mathcal{C}[C_1[P]; M]$, will have the form $C[N] + M$. Moreover, any such interpretation will satisfy the theory $C_1[P]^+_M$ iff it satisfies the theory $C_1[P]$. Indeed, a rule in $C_1[P]$ with a negative literal $\neg\alpha$, such that $\alpha \in M$, is automatically satisfied by $C[N]+M$. On the other hand, a rule in $C_1[P]$

$$\mathrm{C}\alpha_1 \wedge \cdots \wedge \mathrm{C}\alpha_n \wedge \neg\beta_1 \wedge \cdots \wedge \beta_m \Rightarrow \mathrm{C}\gamma$$

in which no negated literal β_i belongs to M, is satisfied by $C[N] + M$ iff a corresponding rule

$$\mathrm{C}\alpha_1 \wedge \cdots \wedge \mathrm{C}\alpha_n \Rightarrow \mathrm{C}\gamma$$

in $C_1[P]^+_M$ is. So both classes $\mathcal{C}[C_1[P]^+_M; M]$ and $\mathcal{C}[C_1[P]; M]$ contain the same models, and thus, one is perfectly coherent only if the other is. Since $\mathcal{C}_M$ in the lemma is simply an abbreviation for $\mathcal{C}[C_1[P]; M]$, the lemma is thus proven. ∎

Theorem 5.1 *Let P be a stratified program. Then M is the canonical model of P if and only if $\mathcal{C}_M$ is the single causally preferred class of $C_1[P]$.*

Proof Let us recall that for a class $\mathcal{C}$ of a theory T, $\Delta[\mathcal{C}]$ denotes the gap of $\mathcal{C}$, and $\Delta^c[\mathcal{C}]$ the explained gap of $\mathcal{C}$. Moreover, for two classes $\mathcal{C}$ and $\mathcal{C}'$ of T, $\mathcal{C}$ is as (causally) preferred as $\mathcal{C}'$ if $\Delta[\mathcal{C}] - \Delta^c[\mathcal{C}] \subseteq \Delta[\mathcal{C}']$, and $\mathcal{C}$ is (causally) preferred to $\mathcal{C}'$ if $\mathcal{C}$ is as preferred as $\mathcal{C}'$ but $\mathcal{C}'$ is not as preferred as $\mathcal{C}$. Likewise, a perfectly coherent class $\mathcal{C}$ of T is a class for which $\Delta[\mathcal{C}] = \Delta^c[\mathcal{C}]$. Now, for a stratified program P, the canonical model M of P and the single stable model of P coincide [Gelfond and Lifschitz, 1988, Van Gelder *et al.*, 1988]. As a result of lemma 5.1 then, M is a canonical model of P iff the class $\mathcal{C}_M$ of models of $C_1[P]$ with an atomic gap M is a perfectly coherent class. Moreover, the class $\mathcal{C}_M$, having an empty unexplained gap, is as preferred as any other class of $C_1[P]$. To prove the theorem, thus, we only need to show that there is no class $\mathcal{C}_{M'}$ of T, with $M' \neq M$,

which is as preferred as $\mathcal{C}_M$. We will write the relation $\alpha_1 < \alpha_2$, for two atoms α_1 and α_2 with predicates p_1 and p_2, respectively, when the dependency graph of P contains a path connecting p_1 to p_2 which includes a negative link.[5] Since P is stratified, the relation '$<$' must be a strict partial order. So, let α be a *minimal* element in $M' - M$ relative to the order '$<$'. For the class $\mathcal{C}'_M$ to be as preferred as $\mathcal{C}_M$, the atom α must have an explanation in $\mathcal{C}_{M'}$. We show that there must be a set of atoms $\alpha_1, \ldots, \alpha_n$ not in M', such that $C_1[P], \neg\alpha_1, \ldots, \neg\alpha_n \vdash \mathrm{C}\alpha$, and $\alpha_i < \alpha$, for $i = 1, \ldots, n$. Indeed, by arguments similar to the those in the proof of lemma 5.1, if $\mathrm{C}\alpha$ holds in the class $\mathcal{C}'_M$, $\mathrm{C}\alpha$ must also be a logical consequence of the positive causal theory $C_1[P]^+_{M'}$. In that case there must be a proof for $\mathrm{C}\alpha$ in $C_1[P]^+_{M'}$ involving only rules with $\mathrm{C}\alpha$ in their heads, or rules that *precede* some of these; where a rule with head H precedes the rules with H in its body, and any other rules the latter rules precede. Thus, the literals $\neg\alpha_1, \ldots, \neg\alpha_n$ can be selected, for instance, as the literals which removed from $C_1[P]_M$, span a set of the rules in $C_1[P]^+_{M'}$ which legitimize one such proof.

Now, one of the atoms α_i must belong to M; otherwise, α would belong to M as well. So, let α_i be an atom in M. Since $\mathcal{C}_M$ is a perfectly coherent class, such an atom must also have an explanation in $\mathcal{C}_M$. By arguments similar to the ones above, there must be a non-empty set of atoms $\alpha'_1, \ldots, \alpha'_{m_i}$ not in M, such that $C_1[P], \neg\alpha'_1, \ldots, \neg\alpha'_{m_i} \vdash \mathrm{C}\alpha_i$, and $\alpha'_j < \alpha_i$ holds, for $j = 1, \ldots, m_i$. Furthermore, in such case, on such atom α'_j, $1 \leq j \leq m_i$ must now belong to M'; for, otherwise, the atom α_i could not be false in M'. This, however, contradicts the minimality of α, as both $\alpha'_j < \alpha_i$ and $\alpha_i < \alpha$ hold. Thus, there cannot be a class $\mathcal{C}'_M$ as preferred as $\mathcal{C}_M$ in $C_1[P]$, and thus $\mathcal{C}_M$ is the single causally preferred class. ∎

Theorem 5.2 *For a stratified program P, there is a single induced causal model which is identical to the canonical model of P.*

Proof Let $\mathcal{C}[T; M]$ stand for a class of T with an atomic gap M. We have shown above that $\mathcal{C}[C_1[P]; M]$ is a perfectly coherent class of the causal theory $C_1[P]$, for a stratified program P with a canonical model M. Moreover, since the set of non-causal atoms satisfied by any model in $\mathcal{C}[C_1[P]; M]$ is identical to M, every interpretation in $\mathcal{C}[C_1[P]; M]$ is also a model of $C_2[P]$. Furthermore, since every model of $C_2[P]$ is a model of $C_1[P]$ as well, the class $\mathcal{C}[C_2[P]; M]$ contains the same models as the class $\mathcal{C}[C_1[P]; M]$, and therefore, $\mathcal{C}[C_2[P]; M]$ is perfectly coherent, and thus, a preferred class of $C_2[P]$. Moreover, it has to be unique preferred; for if

[5] See [Apt *et al.*, 1987], for the relevant terminology.

$C[C_2[P]; M']$, with $M' \neq M$, is as preferred as $C[C_2[P]; M]$, so $C[C_1[P]; M']$ should be as preferred as $C[C_1[P]; M]$. Indeed, an atom α in $M' - M$, is explained in class $C[C_2[P]; M']$ iff $C\alpha$ is a logical consequence of the theory $C_2[P]^+_{M'}$, and since the logical consequences of $C_2[P]^+_{M'}$ and $C_1[P]^+_{M'}$ are identical, α will also be explained in the class $C[C_1[P]; M']$. Thus, for a stratified program P, $C[C_2[P]; M]$ is the single causally preferred class iff M is the canonical model of P, and thus, M is the single induced causal model of P. ∎

Theorem 5.3 *Let P be an acyclic program. Then the class C_M, where M is the canonical model of P, is the unique causally preferred class of the theories $C_1[P]$, $C_2[P]$ and $C_3[P]$.*

Proof Since an acyclic program is a stratified program, in light of the results above, all we need to show is that C_M is the single preferred class of $C_3[P]$. Furthermore, recall that if

$$\gamma \leftarrow \alpha_1, \ldots, \alpha_n, \neg\beta_1, \ldots, \neg\beta_m$$

is a rule in P,

$$C\alpha_1 \wedge \ldots \wedge C\alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow C\gamma$$

will be the corresponding rule in $C_1[P]$, and

$$\alpha_1 \wedge \ldots \wedge \alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow C\gamma$$

will be the corresponding rule in $C_3[P]$. Thus, models of $C_3[P]$ are models of $C_1[P]$ and, if M is a model of P (without causal atoms), $C[M] + M$ will be a model of $C_3[P]$. So the collection of models of $C_3[P]$ with a gap M is not empty, and since, they also belong to the perfectly coherent class $C[C_1[P]; M]$, all support the truth of the causal atoms in $C[M]$. Thus, $C_M = C[C_3[P]; M]$ is a perfectly coherent class, and thus, a preferred class of $C_3[P]$. We need to show then, there is no class $C'_M = C[C_3[P]; M']$, with $M' \neq M$, as preferred as C_M. So, let us assume otherwise that C'_M is as preferred as C_M, and let us write $\alpha_1 < \alpha_2$ for two atoms α_1 and α_2 with predicates p_1 and p_2 connected by a (non necessarily negative) directed path in the dependency graph of P. Since the program is acyclic, it is then possible to select an atom α in $M' - M$, which is minimal relative to such order. Moreover, since the class C'_M is as preferred as the class C_M, it follows then, that C'_M must explain α. Namely, $C_3[P]$ must contain a rule:

$$\alpha_1 \wedge \ldots \wedge \alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow \mathrm{C}\alpha$$

such that every positive literal α_i and negative literal $\neg\beta_j$ hold in $\mathcal{C}'_M$. Indeed, since $\mathcal{C}'_M$ must be a minimal class, each non-causal atom holds in a model in $\mathcal{C}'_M$ if and only if it belongs to M'. Furthermore, since $\alpha_i < \alpha$, for $i = 1, \ldots, n$, and α is a *minimal* atom in $M' - M$, it follows that every such positive antecedent α_i of α must also belong to M. Then, since α does not belong to M, one of the atoms β_i, $1 \leq i \leq m$, must belong to $M - M'$. So let α' be the minimal element in $M - M'$, such that $\alpha' < \beta_i$. Such atom α' must then be explained in the class $\mathcal{C}_M$, and therefore, $C_3[P]$ must contain a rule:

$$\alpha'_1 \wedge \ldots \wedge \alpha'_{n_i} \wedge \neg\beta'_1 \wedge \ldots \wedge \neg\beta'_{m_i} \Rightarrow \mathrm{C}\alpha'$$

in which every antecedent α'_i and $\neg\beta'_j$ holds in $\mathcal{C}_M$. Furthermore, no β'_j, $1 \leq j \leq m_i$, may belong to M' since $\beta'_j \notin M$ and $\beta'_j < \alpha' < \beta_i < \alpha$. On the other hand, every atom α'_k must also belong to M', given the minimality of α'. So, every antecedent α'_i and $\neg\beta'_j$ of α' holds in every model in the class $\mathcal{C}'_M$, contradicting the assumption that α' does not belong to M'. Thus, there cannot be a second causally preferred class $\mathcal{C}'_M$ for a causal theory $C_3[P]$ for an acyclic program P, and thus, $\mathcal{C}_M$ is the single preferred class of $C_3[P]$. ∎

Bibliography

[Adams, 1966] E. Adams. Probability and the logic of conditionals. In J. Hintikka and P. Suppes, editors, *Aspects of Inductive Logic*. North Holland Publishing Company, Amsterdam, 1966.

[Adams, 1975] E. Adams. *The Logic of Conditionals*. D. Reiter, Dordrecht, 1975.

[Adams, 1978] E. Adams. A note comparing probabilistic and modal logics of conditionals. *Theoria*, 43:186–194, 1978.

[Allen, 1984] J. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154, 1984.

[Apt *et al.*, 1987] K. Apt, H. Blair, and A. Walker. Towards a theory of declarative knowledge. In J. Minker, editor, *Foundations of Deductive Databases and Logic Programming*, pages 89–148. Morgan Kaufmann, Los Altos, CA, 1987.

[Bacchus, 1989] F. Bacchus. A modest, but semantically well founded, inheritance reasoner. *Proceedings IJCAI-89*, pages 1104–1109, Detroit, MI., 1989.

[Baker and Ginsberg, 1989] A. Baker and M. Ginsberg. A theorem prover for prioritized circumscription. *Proceedings IJCAI-89*, pages 463–467, Detroit, MI., 1989.

[Bossu and Siegel, 1985] G. Bossu and P. Siegel. Saturation, non-monotonic reasoning and the closed-world assumption. *Artificial Intelligence*, 25:13–63, 1985.

[Brachman and Schmolze, 1985] R. Brachman and J. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9:171–216, 1985.

[Chang and Lee, 1973] C. Chang and R. Lee. *Symbolic Logic and Mechanical Theorem Proving*. Academic Press, New York, 1973.

[Charniak and McDermott, 1985] E. Charniak and D. McDermott. *Introduction to Artificial Intelligence.* Addison Wesley, Reading, MA., 1985.

[Clark, 1978] K. Clark. Negation as failure. In H. Gallaire and J. Minker, editors, *Logic and Data Bases*, pages 293–322. Plenum Press, New York, 1978.

[Console *et al.*, 1989] L. Console, D. Dupre, and P. Torasso. A theory of diagnosis for incomplete causal models. *Proceedings IJCAI-89*, pages 1311–1317, Detroit, Michigan, 1989.

[de Kleer, 1986] J. de Kleer. An assumption-based truth maintenance system. *Artificial Intelligence*, 28:280–297, 1986.

[Dean and Boddy, 1987] T. Dean and M. Boddy. Incremental causal reasoning. *Proceedings AAAI-87*, pages 196–201, Seattle, WA., 1987.

[Dean and McDermott, 1987] T. Dean and D. McDermott. Temporal data base management. *Artificial Intelligence*, 32:1–55, 1987.

[Delgrande, 1987] J. Delgrande. An approach to default reasoning based on a first-order conditional logic. *Proceedings AAAI-87*, pages 340–345, Seattle, 1987.

[Doyle, 1979] J. Doyle. A truth maintenance system. *Artificial Intelligence*, 12:231–272, 1979.

[Doyle, 1985] J. Doyle. Expert systems and the "myth" of symbolic reasoning. *IEEE Transactions on Software Engineering*, 11:1386–1390, 1985.

[Dyer, 1983] M. Dyer. *In Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension.* MIT Press, Cambridge, MA., 1983.

[Elkan, 1988] C. Elkan. A rational reconstruction of nonmonotonic TMSs. Technical report, Cornell University, 1988.

[Etherington and Reiter, 1983] D. Etherington and R. Reiter. On inheritance hierarchies with exceptions. *Proceedings AAAI-83*, pages 104–108, Washington, D.C., 1983.

[Etherington *et al.*, 1985] D. Etherington, R. Mercer, and R. Reiter. On the adequacy of predicate circumscription for closed-world reasoning. *Computational Intelligence*, 1:11–15, 1985.

[Etherington, 1988] D. Etherington. *Reasoning with Incomplete Information.* Pitman, London, 1988.

[Fahlman, 1979] S. Fahlman. *NETL: A System for Representing and Using Real-World Knowledge.* MIT Press, Cambridge, MA., 1979.

[Fine, 1989] K. Fine. The justification of negation as failure. *Proceedings of 8th International Congress of Logic Methodology and Philosophy of Science.* North Holland, 1989.

[Gabbay, 1985] D. Gabbay. Theoretical foundations for non-monotonic reasoning in expert systems. In K. R. Apt, editor, *Logics and Models of Concurrent Systems*, pages 439–457. Springer-Verlag, Heilderberg, 1985.

[Gardenfors, 1988] P. Gardenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* MIT Press, Cambridge, MA., 1988.

[Geffner and Pearl, 1987] H. Geffner and Judea Pearl. Sound defeasible inference. Technical Report TR-94, Cognitive Systems Laboratory, UCLA, Los Angeles, CA., August 1987. Revised version to appear as "A Framework for Reasoning with Defaults", in *Knowledge Representation and Defeasible Inference*, H. Kyburg, R. Loui and G. Carlson (Eds), Kluwer, 1989.

[Geffner and Verma, 1989] H. Geffner and T. Verma. Inheritance = Chaining + Defeat. Technical Report TR-129, Cognitive Systems Lab., UCLA, Los Angeles, CA., 1989. Condensed version in *Methodologies for Intelligent Systems 4,* Z. Ras and L. Saitta (Eds), North Holland, 1989.

[Geffner, 1988] H. Geffner. On the logic of defaults. *Proceedings AAAI-88*, pages 449–454, St. Paul, MN, 1988.

[Geffner, 1989] H. Geffner. Default reasoning, minimality and coherence. *Proceedings of the First International Conference on Principle of Knowledge Representation and Reasoning*, pages 137–148, Toronto, Ontario, 1989.

[Gelfond and Lifschitz, 1988] M. Gelfond and V. Lifschitz. The stable model semantics for logic programming. *Proceedings 1988 Symposium on Logic Programming*, pages 1070–1080, Cambridge, Mass., 1988. MIT Press.

[Gelfond and Przymusinska, 1986] M. Gelfond and H. Przymusinska. Negation as failure: careful closure procedure. *Artificial Intelligence*, 30:273–287, 1986.

[Gelfond and Przymusinska, 1989] M. Gelfond and H. Przymusinska. Inheritance reasoning in autoepistemic logic. Technical report, Computer Science Department, University of Texas at El Paso, El Paso, Texas, 1989.

[Gelfond, 1987] M. Gelfond. On stratified autoepistemic theories. *Proceedings AAAI-87*, pages 207–211, Seattle, Washington, 1987.

[Gelfond, 1989] M. Gelfond. Autoepistemic logic and formalization of commonsense reasoning. a preliminary report. In M. Reinfrank *et al.*, editor, *Proceedings of the Second International Workshop on Non-Monotonic Reasoning*, pages 177–186, Berlin, Germany, 1989. Springer Lecture Notes on Computer Science.

[Genesereth and Nilsson, 1987] M. Genesereth and N. Nilsson. *Logical Foundations of Artificial Intelligence.* Morgan Kaufmann, Los Altos, CA., 1987.

[Ginsberg and Smith, 1988] M. Ginsberg and D. Smith. Reasoning about action i: A possible worlds approach. *Artificial Intelligence*, 35:165–195, 1988.

[Ginsberg, 1987] M. Ginsberg, editor. *Readings in Nonmonotonic Reasoning.* Morgan Kaufmann, Los Altos, CA., 1987.

[Ginsberg, 1988] M. Ginsberg. Multivalued logics: A uniform approach to reasoning in artificial intelligence. *Computational Intelligence*, 4:265–316, 1988.

[Ginsberg, 1989] M. Ginsberg. A circumscriptive theorem prover. *Artificial Intelligence*, 39:209–230, 1989.

[Glymour and Thomason, 1984] C. Glymour and R. Thomason. Default reasoning and the logic of theory perturbation. *Proceedings Mon-Monotonic Reasoning Workshop*, pages 93–102, New Paltz, 1984.

[Goldszmidt and Pearl, 1989] M. Goldszmidt and J. Pearl. Deciding consistency of databases containing defeasible and strict information. *Proceedings Workshop on Uncertainty in AI*, 1989.

[Goodman, 1955] N. Goodman. *Fact, Fiction and Forecast.* Harvard University Press, Cambridge, MA., 1955.

[Grosof, 1988] B. Grosof. Non-monotonicity in probabilistic reasoning. In J. Lemmer and L. Kanal, editors, *Uncertainty in Artificial Intelligence 2*, pages 237–249. Elsevier Science Publishers, 1988.

[Hanks and McDermott, 1985] S. Hanks and D. McDermott. Temporal reasoning and default logics. Technical report, Department of Computer Science, Yale University, 1985.

[Hanks and McDermott, 1986] S. Hanks and D. McDermott. Default reasoning, non-monotonic logics, and the frame problem. *Proceedings AAAI-86*, pages 328–333, Philadelphia, 1986.

[Hanks and McDermott, 1987] S. Hanks and D. McDermott. Non-monotonic logics and temporal projection. *Artificial Intelligence*, 33:379–412, 1987.

[Harman, 1986] G. Harman. *Change in View.* MIT Press, Cambridge, Mass., 1986.

[Haugh, 1987] B. Haugh. Simple causal minimizations for temporal persistence and projection. *Proceedings of the AAAI-87*, pages 218–223, Seattle, Washington, 1987.

[Hewitt, 1972] C. Hewitt. Description and theoretical analysis of planner: a language for proving theorems and manipulating models in a robot. Technical Report TR-258, MIT, AI Lab., Cambridge, Mass., 1972.

[Horty *et al.*, 1987] J. Horty, R. Thomason, and D. Touretzky. A skeptical theory of inheritance. *Proceedings AAAI-87*, pages 358–363, Seattle, Washington, 1987.

[Hughes and Cresswell, 1968] G. Hughes and M. Cresswell. *An Introduction to Modal Logic.* Methuen and Co. LTD, London, Great Britain, 1968.

[Kautz and Selman, 1989] H. Kautz and B. Selman. Hard problems for simple default logics. *Proceedings of the First International Conference on Principle of Knowledge Representation and Reasoning*, pages 189–197, Toronto, Ontario, 1989.

[Kautz, 1987] H. Kautz. *A Formal Theory of Plan Recognition.* PhD thesis, University of Rochester, Rochester, N.Y., May 1987 1987.

[Kolodner, 1984] J. Kolodner. *Retrieval and Organizational Strategies in Conceptual Memory.* Lawrence Erlbaum Associates, Hillsdale, N.J., 1984.

[Konolige and Myers, 1989] K. Konolige and K. Myers. Representing defaults with epistemic concepts. *Computational Intelligence*, 5:32–44, 1989.

[Konolige, 1988] K. Konolige. On the relation between default logic and autoepistemic logic. *Artificial Intelligence*, 35:343–382, 1988.

[Kowalski and Sergot, 1986] R. Kowalski and M. Sergot. A logic-based calculus of events. *New Generation Computing*, 4:67–95, 1986.

[Kowalski, 1979] R. Kowalski. Algorithm = Logic + Control. *Communications of the ACM*, 22:424–436, 1979.

[Kraus *et al.*, 1988] S. Kraus, D. Lehmann, and M. Magidor. Preferential models and cumulative logics. Technical report, Dept. of Computer Science, Hebrew University, Jerusalem 91904, Israel, August 1988.

[Krishnaprasad *et al.*, 1989] T. Krishnaprasad, M. Kiefer, and D. Warren. On the circumscriptive semantics of inheritance networks. In Z. Ras and L. Saitta, editors, *Methodologies for Intelligent Systems 4*. North Holland, New York, N.Y., 1989.

[Kyburg, 1983] H. Kyburg. The reference class. *Philosophy of Science*, 50:374–397, 1983.

[Lehmann and Magidor, 1988] D. Lehmann and M. Magidor. Rational logics and their models: a study in cumulative logic. Technical report, Dept. of Computer Science, Hebrew University, Jerusalem 91904, Israel, November 1988.

[Lehmann, 1989] D. Lehmann. What does a conditional knowledge base entail? *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 212–222, Toronto, Ontario, 1989. Morgan Kaufmann.

[Levesque and Brachman, 1987] H. Levesque and R. Brachman. Expressiveness and tractability in knowledge representation and reasoning. *Computational Intelligence*, 3:78–93, 1987.

[Levesque, 1987] H. Levesque. All I know: An abridged report. *Proceedings AAAI-87*, pages 426–431, Seattle, WA., 1987.

[Lewis, 1973] D. Lewis. *Counterfactuals*. Harvard University Press, Cambridge, MA, 1973.

[Lifschitz, 1985] V. Lifschitz. Computing circumscription. *Proceedings IJCAI-85*, pages 121–127, Los Angeles, CA, 1985.

[Lifschitz, 1987] V. Lifschitz. Formal theories of action. *Proceedings of the 1987 Workshop on the Frame Problem in AI*, pages 35–57, Kansas, 1987.

[Lifschitz, 1988a] V. Lifschitz. Circumscriptive theories: a logic-based framework for knowledge representation. *Journal of Philosophical Logic*, 17:391–441, 1988.

[Lifschitz, 1988b] V. Lifschitz. On the declarative semantics of logic programs. In J. Minker, editor, *Foundations of Deductive Databases and Logic Programming*, pages 177–192. Morgan Kaufmann, Los Altos, CA., 1988.

[Lloyd, 1984] J. Lloyd. *Foundations of Logic Programming.* Springer-Verlag, New York, 1984.

[Loui, 1987a] R. Loui. Defeat among arguments: A system of defeasible inference. *Computational Intelligence*, 1987.

[Loui, 1987b] R. Loui. Real rules of inference. *Communication and Cognition*, 1987.

[Makinson, 1989] D. Makinson. General theory of cumulative inference. In M. Reinfrank *et al.*, editor, *Proceedings of the Second International Workshop on Non-Monotonic Reasoning*, pages 1–18, Berlin, Germany, 1989. Springer Lecture Notes on Computer Science.

[Maloney, 1989] J. Maloney. In praise of narrow minds: the frame problem. In J. Fetzer, editor, *Aspects of Artificial Intelligence*, pages 55–80. Kluwer Academic Publishers, 1989.

[Marek, 1986] W. Marek. Stable theories in autoepistemic logic. Technical report, University of Kentucky, Lexignton, KY., 1986.

[McCarthy and Hayes, 1969] J. McCarthy and P. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Mitchie, editors, *Machine Intelligence 4*, pages 463–502. American Elsevier, New York, 1969.

[McCarthy, 1968] J. McCarthy. Programs with commonsense. In M. Minsky, editor, *Semantic Information Processing.* MIT Press, Cambridge, MA., 1968.

[McCarthy, 1980] J. McCarthy. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980.

[McCarthy, 1986] J. McCarthy. Applications of circumscription to formalizing commonsense knowledge. *Artificial Intelligence*, 28:89–116, 1986.

[McCarthy, 1987] J. McCarthy. Generality in artificial intelligence. *Communications of the ACM*, 30, 1987.

[McClelland and Rumelhart, 1986] J. McClelland and D. Rumelhart, editors. *Parallel Distributed Processing: Explorations into the Microestructure of Cognition*, volume 2. MIT Press, Cambridge, MA., 1986.

[McDermott and Doyle, 1980] D. McDermott and J. Doyle. Non-monotonic logic I. *Artificial Intelligence*, 13:41–72, 1980.

[McDermott, 1982] D. McDermott. A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6:101–155, 1982.

[McDermott, 1987] D. McDermott. Logic, problem solving and deduction. *Annual Review of Computer Science*, 2:187–229, 1987.

[Moore, 1985a] R. Moore. A formal theory of knowledge and action. In J. Hobbs and R. Moore, editors, *Formal Theories of the Commonsense World.* Ablex Publishing Co., Norwood, N.J., 1985.

[Moore, 1985b] R. Moore. Semantical considerations on non-monotonic logics. *Artificial Intelligence*, 25:75–94, 1985.

[Morgenstern and Stein, 1988] L. Morgenstern and L. Stein. Why things go wrong: a formal theory of causal reasoning. *Proceedings AAAI-88*, St. Paul, Minnesota, 1988.

[Myers and Smith, 1988] K. Myers and D. Smith. The persistence of derived information. *Proceedings AAAI-88*, pages 496–500, St. Paul, Minnesota, 1988.

[Neufeld and Poole, 1988] E. Neufeld and D. Poole. Probabilistic semantics and defaults. *Proceedings 4th AAAI Workshop on Uncertainty in AI*, pages 275–281, Minneapolis, MN., 1988.

[Nilsson, 1986] N. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28:71–88, 1986.

[Nute, 1984] D. Nute. Conditional logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, pages 387–439. D. Reidel, Dordrecht, 1984.

[Nute, 1986] D. Nute. LDR: a logic for defeasible reasoning. Technical Report ACMC Research Report 01-0013, University of Georgia, Athens, 1986.

[Pearl and Geffner, 1988] J. Pearl and H. Geffner. Probabilistic semantics for a subset of default reasoning. Technical report, UCLA, Los Angeles, CA., 1988.

[Pearl, 1988a] J. Pearl. Embracing causality in default reasoning. *Artificial Intelligence*, 35:259–271, 1988.

[Pearl, 1988b] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, Los Altos, CA., 1988.

[Pearl, 1989a] J. Pearl. Probabilistic semantics for nonmonotonic reasoning: A survey. *Proceedings of the First Int. Conf. on Principles of Knowledge Representation and Reasoning*, pages 505–516, Toronto, Canada, 1989.

[Pearl, 1989b] J. Pearl. System Z: A natural ordering of defaults with tractable applications to non-monotonic reasoning. Technical report, UCLA, 1989.

[Peirce, 1955] C. Peirce. *Abduction and Induction.* Dover, New York, 1955.

[Pitts, 1988] J. Pitts, editor. *Theories of Explanation.* Oxford University Press, New York, N. Y., 1988.

[Pollock, 1987] J. Pollock. Defeasible reasoning. *Cognitive Science*, 11:481–518, 1987.

[Pollock, 1988] J. Pollock. Defeasible reasoning and the statistical syllogism. Unpublished manuscript, 1988.

[Poole, 1985] D. Poole. On the comparison of theories: Preferring the most specific explanation. *Proceedings of IJCAI-85*, pages 144–147, Los Angeles, 1985.

[Poole, 1987] D. Poole. Defaults and conjectures: hypothetical reasoning for explanation and prediction. Technical Report CS-87-4, University of Waterloo, 1987.

[Poole, 1988] D. Poole. A logical framework for default reasoning. *Artificial Intelligence*, 36:27–47, 1988.

[Poole, 1989] D. Poole. Normality and faults in logic-based diagnosis. *Proceedings IJCAI-89*, pages 1304–1310, Detroit, Michigan, 1989.

[Przymusinska and Przymusinski, 1988] H. Przymusinska and T. Przymusinski. Weakly perfect model semantics for logic programs. In R. Kowalski and K. Bowen, editors, *Proceedings of the Fifth Logic Programming Symposium*, pages 1106–1122, Cambridge, Mass., 1988. MIT Press.

[Przymusinska and Przymusinski, 1989] H. Przymusinska and T. Przymusinski. Semantic issues in deductive databases and logic programs. In A. Banerji, editor, *Sourcebook on the Formal Approaches in Artificial Intelligence.* North Holland, Amsterdam, 1989.

[Przymusinski, 1987] T. Przymusinski. On the declarative semantics of stratified deductive databases and logic programs. In J. Minker, editor, *Foundations of Deductive Databases and Logic Programming*, pages 193–216. Morgan Kaufmann, Los Altos, CA, 1987.

[Przymusinski, 1988] T. Przymusinski. On the relationship between non-monotonic reasoning and logic programming. *Proceedings AAAI-88*, pages 444–448, St. Paul, Minnesota, 1988.

[Przymusinski, 1989] T. Przymusinski. Three-valued non-monotonic formalisms and logic programming. *Proceedings of the First International Conference on Principle of Knowledge Representation and Reasoning*, pages 341–340, Toronto, Ontario, 1989.

[Reggia *et al.*, 1985] J. Reggia, D. Nau, P. Wang, and Y. Peng. A formal model of abductive inference. *Information Sciences*, 37:227–285, 1985.

[Reinfrank *et al.*, 1989] M. Reinfrank, O. Dressler, and G. Brewka. On the relation between truth maitenance systems and autoepistemic logic. *Proceedings IJCAI-89*, pages 1206–1212, Detroit, MI., 1989.

[Reiter and Criscuolo, 1983] R. Reiter and G. Criscuolo. Some representational issues in default reasoning. *Int. J. of Computers and Mathematics*, 9:1–13, 1983.

[Reiter and de Kleer, 1987] R. Reiter and J. de Kleer. Foundations of assumption-based truth maintenance systems: a preliminary report. *Proceedings AAAI-87*, pages 183–188, Seattle, WA, 1987.

[Reiter, 1980] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 12:81–132, 1980.

[Reiter, 1984] R. Reiter. Towards a logical reconstruction of relational database theory. In M. Brodie, J. Mylopoulos, and J. W. Schmidt, editors, *On Conceptual Modelling*, pages 163–189. Springer-Verlag, New York, 1984.

[Reiter, 1987a] R. Reiter. Nonmonotonic reasoning. *Annual Review of Computer Science*, 2:147–186, 1987.

[Reiter, 1987b] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.

[Rousell, 1975] P. Rousell. Prolog, manuel de reference et d'utilisation. Technical report, Groupe d'Intelligence Artificielle, U.E.R. de Marseille, France, 1975.

[Rumelhart and McClelland, 1986] D. Rumelhart and J. McClelland, editors. *Parallel Distributed Processing: Explorations into the Microestructure of Cognition*, volume 1. MIT Press, Cambridge, MA., 1986.

[Sandewal, 1988] E. Sandewal. An approach to non-monotonic entailment. In Z. Ras and L. Saitta, editors, *Methodologies for Intelligent Systems 3*, pages 391–397. North Holland, New York, N.Y., 1988.

[Schank and Abelson, 1977] R. Schank and R. Abelson. *Scripts, Plans, Goals and Understanding.* Laurence Erlbaum Associates, Hillsdale, N.J., 1977.

[Selman and Levesque, 1989] B. Selman and H. Levesque. The tractability of path-based inheritance. *Proceedings IJCAI-89*, pages 1140–1145, Detroit, MI., 1989.

[Shafer, 1976] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, 1976.

[Shepherson, 1987] J. Shepherson. Negation in logic programming. In J. Minker, editor, *Foundations of Deductive Databases and Logic Programming*, pages 19–88. Morgan Kaufmann, Los Altos, CA, 1987.

[Shoham, 1986] Y. Shoham. Chronological ignorance: time, non-monotonicity, necessity and causal theories. *Proceedings AAAI-86*, pages 389–393, Philadelphia, 1986.

[Shoham, 1987] Y. Shoham. Temporal logics in AI: Semantical and ontological considerations. *Artificial Intelligence*, 33:89–104, 1987.

[Shoham, 1988] Y. Shoham. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence.* MIT Press, Cambridge, Mass., 1988.

[Sowa, 1984] J. Sowa. *Conceptual Structures: Information Proceeding in Mind and Machine.* Addison-Wesley, Reading, MA., 1984.

[Spohn, 1988] W. Spohn. A general non-probabilistic theory of inductive reasoning. *Proceedings 4th Workshop on Uncertainty*, pages 315–322, St. Paul, 1988.

[Stallman and Sussman, 1977] R. Stallman and G. Sussman. Forward reasoning and dependency-directed backtracking in a system for computed aided circuit analysis. *Artificial Intelligence*, 9:135–196, 1977.

[Touretzky *et al.*, 1987] D. Touretzky, J. Horty, and R. Thomason. A clash of intuitions: The current state of non-monotonic multiple inheritance systems. *Proceedings of IJCAI-87*, pages 476–482, Milano, Italy, 1987.

[Touretzky, 1986] D. Touretzky. *The Mathematics of Inheritance Systems.* Pitman, London, 1986.

[Ullman, 1982] J. Ullman. *Principles of Database Systems.* Computer Science Press, Rockville, Maryland, 1982.

[Van Gelder *et al.*, 1988] A. Van Gelder, K. Ross, and J. S. Schlipf. Unfounded sets and well-founded semantics for general logic programs. *Proceedings Seventh Symp. on Principles of Database Systems*, pages 221–230, 1988.

[Wellman, 1987] M. Wellman. Probabilistics semantics for qualitative influences. *Proceedings AAAI-87*, pages 660–664, Seattle, WA, 1987.

[Winslett, 1988] M. Winslett. Reasoning about action using a possible models approach. *Proceedings AAAI-88*, pages 89–93, St. Paul, Minnesota, 1988.

[Yager *et al.*, 1987] R. Yager, S. Ovchinnikov, S. Tong, and H. Nguyen, editors. *Fuzzy Sets and Applications: Selected papers by L. A. Zadeh.* Wiley, New York, 1987.