COGNITIVE SCIENCE **11**, 481–518 (1987)

Defeasible Reasoning

JOHN L. POLLOCK
University of Arizona

What philosophers call defeasible reasoning is roughly the same as nonmonotonic reasoning in AI. Some brief remarks are made about the nature of reasoning and the relationship between work in epistemology, AI, and cognitive psychology. This is followed by a general description of human rational architecture. This description has the consequence that defeasible reasoning has a more complicated structure than has generally been recognized in AI. We define a proposition to be *warranted* if it would be believed by an ideal reasoner. A general theory of warrant, based on defeasible reasons, is developed. This theory is then used as a guide in the construction of a theory of defeasible reasoning, and a computer program implementing that theory. The theory constructed deals with only a subset of defeasible reasoning, but it is an important subset.

PART I

1. INTRODUCTION

A common misconception about reasoning is that reasoning is deducing, and in good reasoning the conclusions follow logically from the premises. It is now generally recognized both in philosophy and in AI that nondeductive reasoning is at least as common as deductive reasoning, and a reasonable epistemology must accommodate both. For instance, inductive reasoning is not deductive, and in perception, when one judges the color of something on the basis of how it looks to him, he is not reasoning deductively. Such reasoning is *defeasible,* in the sense that the premises taken by themselves may justify us in accepting the conclusion, but when additional information is added, that conclusion may no longer be justified. For example, something's looking red to me may justify me in believing that it is red, but if I subsequently learn that the object is illuminated by red lights and I know that that can make things look red when they are not, then I cease to be justified in believing that the object is red.

Although the existence of nondefeasible reasoning is obvious once it is pointed out, its recognition was slow in coming to philosophy, and it has wrought fundamental changes in epistemology. It is interesting that work on machine reasoning in AI has followed a roughly parallel course. What is

Corresondence and requests for reprints should be sent to John L. Pollock, Department of Philosophy, 213 Social Sciences Bldg. #27, The University of Arizona, Tucson, AZ 85721.

called *nonmonotonic reasoning* in AI is much the same thing as defeasible reasoning in philosophy. Nonmonotonic reasoning became the subject of intense interest in AI at almost the same time philosophical theories of defeasible reasoning were first being developed.[1]

Work in philosophy and computer science can be fruitfully integrated in the study of defeasible reasoning. Philosophy comes to the investigation with an extensive and sophisticated background knowledge about various kinds of reasoning. In designing a program for nonmonotonic reasoning, one must first know what it should do, and here the researcher in AI can learn from the philosopher. As I will argue below, current theories of nonmonotonic reasoning coming out of AI are simplistic and overlook much of the fine structure of defeasible reasoning. Conversely, the philosopher can learn much from attempts to implement epistemological theories in concrete programs for machine reasoning. If a theory of reasoning is correct, it must be possible to build a machine that reasons that way. A theory that looks good in the abstract may not work when you write a program to implement it. From the comfort of his armchair, an epistemologist may be able to find a certain number of counterexamples to a false epistemological theory, but my experience has been that when a program is constructed to implement a theory, it will almost invariably work incorrectly at first, and diagnosing the difficulty leads directly to the discovery of counterexamples to the epistemological theory. In effect, the use of computers constitutes a mechanical way of generating counterexamples.

Philosophers know a lot about some aspects of defeasible reasoning. They know about prima facie reasons and defeaters, and they know quite a bit about what prima facie reasons there are. But they do not have a good understanding of precisely how these constituents are used in reasoning. This is analogous to knowing what primitive logical entailments there are, but not knowing the principles for constructing deductive arguments out of those entailments. The purpose of this paper is to investigate the structure of defeasible reasoning. Given an array of defeasible and nondefeasible reasons, how are they to be used in drawing conclusions? A satisfactory theory of defeasible reasoning ought to be sufficiently precise that it can be implemented in a computer program. Constructing such a computer program and seeing that it does the right thing will be a useful test of the theory, and simultaneously a contribution to AI. Thus, the purpose of this paper is to investigate defeasible reasoning from a theoretical point of view and then discuss a computer program that attempts to implement the theory. The results detailed here are only partial. I will construct a theory that appears

[1] The main philosophical work on defeasible reasoning has been done by Roderick Chisholm, 1957, 1966, and 1977; and myself, 1967, 1970, 1974, 1986. The seminal works in AI are probably Doyle (1979), McCarthy (1980), Reiter (1978) and (1980), and McDermott and Doyle (1980).

correct given certain simplifying assumptions, but a fully adequate theory must relax those simplifying assumptions, and that will be the subject of further research.

2. A SKETCH OF HUMAN RATIONAL ARCHITECTURE

I will begin with some general remarks about human rational architecture. This section provides a brief sketch of the more detailed account defended in my (1986). In order to get on with the business of analyzing reasoning, I will assume without further defense a number of conclusions from that book. The reader who is worried about those assumptions should consult the book for further discussion and defense.

Reasoning is guided by rules, and when the reasoning is in accordance with the rules the resultant beliefs are said to be *justified.* Epistemic rules are normative, and that suggests that their formulation is independent of psychology. But I have argued (Pollock, 1986) that that is a mistake. We know how to reason. That is, reasoning is governed by internalized rules. The possession of such internalized rules constitues procedural knowledge. Jointly, these rules comprise a production system. These rules are the rules for "correct reasoning"—the very rules that are the subject of epistemology. One does not always conform to these rules, but this is because the production system for reasoning is embedded in a larger system than can override it. Reasoning can be fruitfully compared to using language. We have procedural knowledge governing the production of grammatical utterances, but our utterances are not always grammatical. This gives rise to the "competence/performance" distinction in linguistics. Precisely the same distinction needs to be made in connection with any procedural knowledge. In particular, it can be made in connection with reasoning, and it is the existence of this distinction that is responsible for the use of normative language in epistemology. What we "should" do is what the rules of our production system tell us to do, but we do not always conform to those rules because they are embedded in a larger system that can override them.

Human reasoning begins from various kinds of input states, the most familiar of which are straightforward perceptual states. These are nondoxastic states. For instance, something can look red to you without your having any belief to that effect. Furthermore, you can reason from the way things look to conclusions about their objective properties without forming intermediate beliefs about your own perceptual states. If you look around, you will form myriad beliefs about your surroundings but few if any beliefs about how things look to you. You do not usually attend to the way things look to you. This is important in the present context because it means that if we are to describe this process in terms of reasoning, then we must acknowledge that reasons need not be beliefs. At the very least, perceptual states can

be reasons. In general, I will call the states from which reasoning begins *foundational states.*

Crudely put, reasoning proceeds in terms of reasons. Reasons are strung together into arguments and in this way the conclusions of the arguments become justified. The general notion of a reason can be defined as follows:

(2.1) Being in states $M_1,\ldots,M_n$ is a *reason* for S to believe Q if and only if it is logically possible for S to be justified in believing Q on the basis of being in states $M_1,\ldots,M_n$.

Usually, reasons are beliefs or sets of beliefs, and in that case, rather than talking about *believing P* being a reason for believing Q, I will say more simply that P is a reason for Q (or more generally, that a finite set $\{P_1,\ldots,P_n\}$ is a reason for Q).

There are two kinds of reasons—*defeasible* and *nondefeasible.* Nondefeasible reasons are those reasons that logically entail their conclusions. For instance, (P & Q) is a nondefeasible reason for P. Such reasons are *conclusive reasons.* Everyone has always recognized the existence of nondefeasible reasons, but defeasible reasons are a relatively new discovery in philosophy, as well as in allied disciplines like AI. Focusing first upon reasons that are beliefs, P is a defeasible reason for Q just in case P is a reason for Q, but adding additional information may destroy the reason connection. Such reasons are called "prima facie reasons." This notion can be defined more precisely as follows:

(2.2) P is a *prima facie reason* for S to believe Q if and only if P is a reason for S to believe Q and there is an R such that R is logically consistent with P but (P & R) is not a reason for S to believe Q.

(To keep this and subsequent definitions simple, I just formulate them for the case of a reason that is a single proposition rather than a set of propositions, but the general case is analogous.) There are lots of rather obvious examples of prima facie reasons. For instance, "X looks red to me" is a reason for me to believe that X is red, but it is defeasible because if I conjoin it with something like "Jones, who is very reliable, told me that X is not really red but just looks that way because of peculiar lighting conditions," the resulting conjunction cannot justify me in believing that X is red and hence is no longer a reason. Similarly, observing lots of ravens and seeing that they are all black may justify me in believing that all ravens are black. But observing a single additional raven that is not black defeats the reasoning. The R's that defeat prima facie reasons are called "defeaters":

(2.3) R is a *defeater* for P as a prima facie reason for Q if and only if P is a reason for S to believe Q and R is logically consistent with P but (P & R) is not a reason for S to believe Q.

It is prima facie reasons and defeaters that are responsible for the nonmonotonic character of human reasoning.

There are two kinds of defeaters for prima facie reasons. "Rebutting defeaters" are reasons for denying the conclusion:

(2.4) R is a *rebutting defeater* for P as a prima facie reason for Q if and only if R is a defeater and R is a reason for believing $\sim Q$.

Rebutting defeaters are reasonably familiar, and they form the basis for most current AI work on nonmonotonic reasoning. But equally important are *undercutting defeaters,* which attack the connection between the reason and the conclusion rather than attacking the conclusion itself. For instance, "X looks red to me" is a prima facie reason for me to believe that X is red. Suppose I discover that X is illuminated by red lights and illumination by red lights often makes things look red when they are not. This is a defeater, but it is not a reason for denying that X is red (red things look red in red light too). Instead, this is a reason for denying that X wouldn't look red to me unless it were red.

(2.5) R is an *undercutting defeater* for P as a prima facie reason for S to believe Q if and only if R is a defeater and R is a reason for denying that P wouldn't be true unless Q were true.

Undercutting defeaters have generally been overlooked both in philosophy and in AI.

In (2.5), "P wouldn't be true unless Q were true" is some kind of conditional. I will symbolize it as "$P \rightarrow Q$." This is clearly not a material conditional, but beyond that it is unclear how it is to be analyzed. Fortunately, that will make no difference to our present concerns.[2]

Defeaters are defeaters by virtue of being reasons for either $\sim Q$ or $\sim(P \rightarrow Q)$. They may be only defeasible reasons for these conclusions, in which case their defeaters are "defeater defeaters." There may similarly be defeater defeater defeaters, and so on.

The concepts of conclusive reasons, prima facie reasons, rebutting defeaters, and undercutting defeaters, provide the building blocks for a theory of reasoning. My main purpose here is to discuss the general logical structure of reasoning, but that can only be done against the background of some specific kinds of reasons. Thus, the next section will sketch some important classes of reasons. These will then be used as examples in constructing a general theory of reasoning.

---

[2] I used to maintain that "$(P \rightarrow Q)$" was analyzable as $(\sim Q > \sim P)$, where '$>$' is the so-called "simple subjunctive". (See Pollock [1976], chapters one and two, for an informal discussion of the simple subjunctive conditional.) Contraposition fails for subjunctive conditionals, so on this analysis, "$(P \rightarrow Q)$" cannot be written more simply as "$P > Q$." However, I am no longer convinced that this analysis is correct.

3. SOME SUBSTANTIVE REASONS

3.1 Deductive Reasons

Human reasoning proceeds in terms of a number of different kinds of reasons. I assume that there is a set of conclusive reasons enabling us to reason deductively, and that these are sufficiently inclusive to enable us to become justified in believing any theorem of the predicate calculus. These will include reasons like the following:

(3.1) (P & Q) is a conclusive reason for P and for Q.

(3.2) {P,Q} is a conclusive reason for (P & Q).

(3.3) {P,(P⊃Q)} is a conclusive reason for Q.

It is of some interest to try to produce a list of reasons that plausibly reproduce this aspect of human rational architecture, but I will not pursue that here.

3.2 Perception

Perception represents the basic source of human knowledge. Nonintellectual mechanisms put us into various perceptual states, and being in those perceptual states constitutes a prima facie reason for conclusions about the world around us. Philosophers have found it useful to adopt a technical "appeared to" terminology for the formulation of perceptual states. For instance, a perceptual state might consist of being appeared to as if there is something red before me. For appropriate choices of P, *being appeared to as if P* is a perceptual state. We need not worry here about how to delineate the range of appropriate P's. Given such a P, we have the following prima facie reason:

(3.4) "I am appeared to as if P" is a prima facie reason for me to believe P.

For instance, "I am appeared to as if there is something red before me" is a prima facie reason for me to believe that there is something red before me.

Any reason for denying P is a rebutting defeater for (3.4). All prima facie reasons are subject to a kind of undercutting defeater that I call *reliability defeaters*. In general, if P is a prima facie reason for Q then discovering that the present circumstances are of some type C under which P's being true is not a reliable indicator of Q's being true constitutes a defeater.[3] Applying this to (3.4):

(3.5) The following is an undercutting defeater for (3.4):
The present circumstances are of some type C such that the conditional probability is low of P's being true given that I am in circumstances of type C and am appeared to as if P.

---

[3] A qualification is required on the circumstance type C. It must be "projectible," in the sense discussed below.

3.3 Memory

A common theme in epistemology and AI has been that we may arrive at a belief through reasoning, and then later reject that belief because we come to reject some other belief used in the reasoning supporting it. However, human beings tend to have difficulties remembering the reasoning supporting a belief. When they first arrive at a belief, they may know what their reasoning was. Later, they may recall the belief and use that for constructing reasons for other beliefs. But at that time they may be unable to recall their reasons for the belief, or they may be able to do so only with great difficulty. Insofar as the reasons are stored at all, they are stored separately from the beliefs. Presumably, human memory is organized in this way for the sake of efficiency. Having arrived at a belief, it is usually safe to assume that the reasons for adopting it were unproblematic, so we are more interested in the belief than the reasons and it is more important to be able to recall the belief easily than it is to be able to recall the reasons. Thus, to facilitate search for appropriate stored beliefs, they are stored separately from their reasons. This, of course, is speculation, but it seems reasonable.

If we cannot reliably recall our reasons for our beliefs, then we cannot reliably update our beliefs when we reject elements of the reasoning that underlay their acquisition. If one originally used P in reasoning to Q, but can no longer remember that fact, then he is not being irrational in retaining Q even though he later rejects P. We must regard his continued belief in Q as justified *until* he discovers that it was originally based upon reasoning that he would now reject. In other words, memory itself provides defeasible justification for remembered beliefs. The mental state consisting of a belief becoming occurrent by virtue of being supplied by memory will be called *recollection*. Then we must adopt the following mnemonic prima facie reason:

(3.6) S's recalling P is a prima facie reason for S to believe P.

There are several undercutting defeaters for this prima facie reason. The simplest is:

(3.7) The following is an undercutting defeater for (3.6):
S now recalls P because he originally believed it on the basis of a set of beliefs one of which is false.

This undercutting defeater must be generalized a bit, but I will not pursue that here.[4]

Human beings can misremember (perhaps not a problem for computers), so we must also have:

(3.8) The following is an undercutting defeater for (3.6):
S did not originally believe P for reasons other than (3.6).

---

[4] For a fuller discussion, see my (1986), pages 46ff.

This is not a complete catalogue of defeaters for the mnemonic prima facie reason, but it will do for now.

3.4 Statistical Syllogism

Perception and memory provide the starting points for reasoning. Higher level reasoning often proceeds probabilistically. Perhaps the simplest kind of probabilistic reasoning is that involved in the statistical syllogism, which can be written roughly as follows:

Most F's are G.
This is an F.

Therefore (prime facie), this is a G.

This can be written more accurately as follows:

(3.9) "prob(Gx/Fx) is high, and Fa" is a prima facie reason for "Ga," the strength of the reason being determined by how high the probability is.

However, Principle (3.9) requires qualification. Without constraints on the properties F and G, it turns out that whenever (3.9) gives us a reason for believing Ga, and $\text{prob}(Gx/Fx) < 1$, we can construct other instances of (3.9) involving larger probabilities that give us an even better reason for believing $\sim Ga$.[5] Of course, this can be iterated to generate an even better reason for Ga, and so on. The net result is that each reason supplied by (3.9) is defeated by a rebutting defeater also of the form of (3.9), and hence (3.9) is useless. To avoid this sort of difficulty, the properties to which (3.9) can appeal must be restricted. Just to have a label for the allowed properties, we will say that they are *projectible*. A correct principle of statistical syllogism can then be formulated as follows:

(3.10) If G is projectible with respect to F then
prob(Gx/Fx) is high, and Fa
is a prima facie reason for "Ga."

There is no good theory in the literature concerning what properties are projectible. Piecemeal results are readily obtainable. For example, the above reasoning is blocked by the fact that disjunctions are not usually projectible (more accurately, the class of projectible properties is not closed under disjunction). But no general theory of projectibility is available at this time.[6]

In using statistical syllogism, we should be constrained to make our inferences on the basis of those probabilities that take into account as much information about a as possible. This is captured by the following undercutting defeater:

---

[5] This is demonstrated in my (1983).

[6] For further discussion of this, see Pollock (in preparation).

(3.11) If G is projectible with respect to (F & H) then
prob(Gx/Fx & Hx)<prob(Gx/Fx) and Ha
is an undercutting defeater for (3.10).

For instance, suppose we know that the probability of a person getting to his destination by driving is .99, but the probability of a person getting to his destination by driving while he is so drunk he cannot stand up is only .5. If we know that Jones is driving home and is so drunk he cannot stand, the first probability gives us a prima facie reason for thinking he will get home. But the second probability gives us an undercutting defeater for that instance, leaving us unjustified in drawing any conclusion about whether Jones will get home.

3.5 Induction

There are several different kinds of inductive reasoning. I will just mention two. In *enumerative induction,* we reason from the fact that all the F's we have observed have been G to the conclusion that all F's are G. Such reasoning is obviously defeasible, because further information can make us withdraw the conclusion without taking back our belief in the original data. Enumerative induction proceeds by the following rule:

(3.12) If F is projectible with respect to G then
X is a set of F's, and all the members of X are G
is a prima facie reason for "All F's are G."

It can be argued that the projectibility constraint here is the same as the constraint on statistical syllogism.[7]

In *statistical induction* we reason from the fact that a certain proportion of all the F's we have observed have been G to the conclusion that the probability of an F being a G is approximately the same as that proportion:

(3.13) If F and "~F" are projectible with respect to G, then
X is a set of F's, and the proportion of members of X that are G is r
is a prima facie reason for "prob(Gx/Fx) is approximately r."

There are various kinds of "fair sample" undercutting defeaters for these prima facie reasons that attack the reasons on the grounds that the set X was somehow inappropriately chosen. The details of these defeaters are complicated, and in some cases, problematic, so I will not go into them here. I have argued elsewhere (1983, and in preparation) that the prima facie reasons involved in induction are not primitive, but rather can be derived from statistical syllogism and a sufficiently strong calculus of probabilities.

The above has been a sketch of some of the more important kinds of prima facie reasons that are involved in human reasoning. I turn next to the

[7] See my (1984) and (in preparation).

question of how such prima facie reasons are used in determining what one should believe.

PART II

4. WARRANT

In constructing a theory of reasoning, it is useful to begin by considering the fiction of an ideal reasoner, or if you like, an ideal intelligent machine with no limits on memory or computational capacity. How should such a reasoner employ reasons and defeaters in deciding what to believe? Let us say that a proposition is *warranted* in an epistemic situation if and only if an ideal reasoner starting from that situation would be justified in believing the proposition. This section is devoted to giving a precise characterization of the set of warranted propositions.

4.1 Linear Arguments

Reasoning proceeds by arguments, and arguments are constructed by starting from perceptual and memory states, moving from them to beliefs, from those beliefs to new beliefs, and so on. What arguments can be constructed depends upon what perceptual and memory states one is in. Let us take these to comprise the *epistemic basis*.

In the simplest case, an argument is a finite sequence of propositions each of which either describes the epistemic basis or is such that there is a set of earlier members of the sequence that constitutes a reason for it. We might call such arguments *linear arguments*. Not all arguments are linear arguments. There are more complicated kinds of "indirect" arguments that involve subsidiary arguments. Examples of indirect argument include conditionalization, reductio ad absurdum, and the like. Indirect arguments proceed by adopting as premises suppositions that have not been established, using those premises to obtain a conclusion, and then "discharging" the premises by using some rule like conditionalization or reductio ad absurdum. Indirect arguments make the theory more complicated. It is best to begin by adopting the fiction that all arguments are linear arguments. An account of reasoning based on this simplifying assumption will be constructed, and then I will consider how it must be modified to accommodate indirect arguments.

Confining our attention to linear arguments, we can take a line of an argument to be an ordered triple $<P,R,\{m,n,\ldots\}>$ where P is a proposition (the proposition *supported* by the line), R is a rule of inference, and $\{m,n,\ldots\}$ is the set of line numbers of the previous lines to which the rule R appeals in justifying the present line. Linear arguments are finite sequences of such lines. If σ is an argument, σ_i is the proposition supported by its ith line. We say that an argument *supports* a conclusion P if and only if P is supported by some line of the argument.

Arguments are constructed in accordance with "rules of inference." These are just rules for argument formation. They are not necessarily rules of deductive inference, but they will usually be analogous to rules of deductive inference. I assume that the rules for linear argument formation include the following:

Rule **F**: *Foundations*
If P expresses a foundation state contained in the epistemic basis, $<P,\mathbf{F},\emptyset>$ can be entered as any line of the argument.

Rule **R**: *Closure under reasons*
If $\{P_1,\ldots,P_n\}$ is a reason for Q and $<P_1,\ldots>,\ldots,<P_n,\ldots>$ occur as lines $i_1,\ldots,i_n$ of an argument, $<Q,R,\{i_1,\ldots,i_n\}>$ can be entered on any later line.

4.2 Ultimately Undefeated Arguments

Warrant is always relative to a set of foundational states (the epistemic basis) that provide the starting points for arguments. In the following, by "argument" I always mean arguments relative to some fixed epistemic basis. Merely having an argument for a proposition does not guarantee that the proposition is warranted, because one might also have arguments for defeaters for some of the steps in the first argument. Iterating the process, one argument might be defeated by a second, but then the second argument could be defeated by a third thus reinstating the first, and so on. A proposition is warranted only if it ultimately emerges from this process undefeated.

We have temporarily adopted the fiction that we only have to contend with linear arguments. With this simplifying assumption we can give a fairly simple characterization of warrant. The defeat of one argument always results from another argument supporting a defeater for some use of rule **R**, so where η is an argument with at least j lines and σ is an argument with at least i lines, let us define:

(4.1) $<\eta,j>$ *defeats* $<\sigma,i>$ if and only if $<\sigma,i>$ is obtained by rule **R** using $\{P_1,\ldots,P_n\}$ as a prima facie reason for Q, and η_j is either $\sim Q$ or $\sim[(P_1 \,\&\, \ldots \,\&\, P_n)\rightarrow Q]$

Let us say that all arguments are *level 0 arguments.* Some level 0 arguments may provide us with defeaters for lines of other level 0 arguments, so let us say that an argument is a *level 1 argument* if and only if no level 0 argument defeats any of its lines at level 1. As there are fewer level 1 arguments than level 0 arguments, fewer propositions will be supported by level 1 arguments than level 0 arguments. In particular, fewer defeaters for level 0 arguments will be supported by level 1 arguments than by level 0 arguments. Thus having moved to level 1 arguments, we may have removed some defeaters and thereby reinstated some level 0 arguments. Let us say that a *level 2 argument* is a level 0 argument having no lines defeated by any level 1 argument. Some level 0 arguments that were defeated at level 1 may be rein-

stated at level 2. Hence, level 2 arguments may support some defeaters that were not supported by level 1 arguments, thus defeating some level 0 arguments that were not defeated by level 1 arguments. Consequently, we take a *level 3 argument* to be any level 0 argument not defeated by level 2 arguments; and so on. In general, a *level $n+1$ argument* is any level 0 argument not defeated by level n arguments. More precisely:

(4.2) σ is a *level $n+1$ argument* if and only if σ is a level 0 argument and there is no level n argument η such that for some i and j, $<\eta,j>$ defeats $<\sigma,i>$.

A given level 0 argument may be defeated and reinstated many times by this alternating process. Only if we eventually reach a point where it stays undefeated can we say that it warrants its conclusion. Let us say that an argument is *ultimately undefeated* if and only if there is some m such that the argument is a level n argument for every $n>m$. On the simplifying assumption that all arguments are linear arguments, it seems that epistemological warrant can then be characterized in terms of arguments that are ultimately undefeated:

(4.3) P is warranted relative to an epistemic basis if and only if P is supported by some ultimately undefeated argument proceeding from that epistemic basis.

To illustrate, consider the three arguments diagrammed in Figure 1. $<\beta,k>$ defeats $<\alpha,i+1>$, and $<\gamma,m>$ defeats $<\beta,j+1>$. It is assumed that nothing defeats γ. Thus γ is ultimately undefeated. Neither α nor β is a level 1 argument, because both are defeated by level 0 arguments. As γ is a level n argument for every n, β is defeated at every level greater than 0, so β is not a level n argument for $n>0$. As a result α is reinstated at level 2, and is a level n argument for every $n>1$. Hence α is ultimately undefeated, and V is warranted.

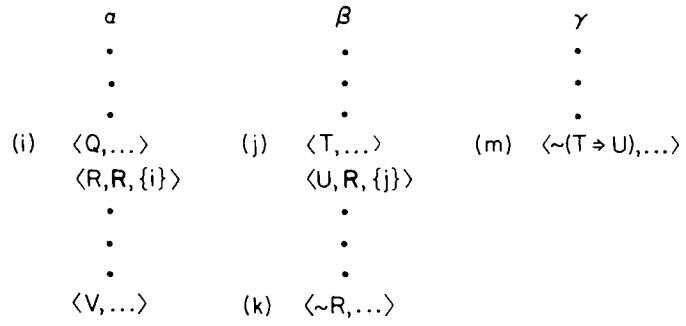

Figure 1

4.3 Collective Defeat

Our analysis of warrant will have to be made slightly more complicated, but in order to appreciate the difficulty we must first note that the analysis entails an important principle of rationality, which I call *the principle of collective defeat*. The following form of the principle follows from the analysis:

(4.4) If we are warranted in believing R and there is a set Γ of propositions such that:
(1) we have equally good defeasible arguments for believing each member of Γ;
(2) for each P in Γ there is a finite subset Γ_P of Γ such that the conjunction of R with the members of Γ_P provides a deductive argument for $\sim P$ that is as strong as our initial argument is for P; and
(3) none of these arguments is defeated except possibly by their interactions with one another;
then none of the propositions in Γ is warranted on the basis of these defeasible arguments.

The proof of this principle is as follows. Suppose we have such a set Γ and proposition R. For each P in Γ, combining the argument supporting R with the arguments supporting the members of Γ_P gives us an argument supporting $\sim P$. Intuitively, we have equally strong support for both P and $\sim P$, and hence we could not reasonably believe either on this basis, that is, neither is warranted. This holds for each P in Γ, so none of them should be warranted. They "collectively defeat one another." And, indeed, this is forthcoming from our analysis of warrant. We have level 0 arguments supporting each P. But these can be combined to generate level 0 arguments that also support rebutting defeaters for the argument for each P. Thus none of these are level 1 arguments. But this means that none of the defeating arguments are level 1 arguments either. Thus all of the arguments are level 2 arguments. But then they fail to be level 3 arguments, and so on. For each even number n, each P is supported by a level n argument, but that argument is not a level $n+1$ argument. Thus, the P's are not supported by ultimately undefeated arguments, and hence are not warranted.

The most common instances of (4.4) occur when Γ is a minimal finite set of propositions deductively inconsistent with R. In that case, for each P in Γ, $\{R\} \cup (\Gamma - \{P\})$ gives us a deductive reason for $\sim p$. Principle (4.4) can be illustrated by an example of this form that has played an important role in the philosophical foundations of probability theory. This is the *lottery paradox* (due to Kyburg, 1961). Suppose we are warranted in believing we have a fair lottery with one million tickets. Let this be R. Then the probability of the ith ticket being drawn in such a lottery is .000001. By statistical syllogism (3.10), this gives us a prima facie reason for believing that the ith ticket will not be drawn. Let the latter be P_i. But we have an analogous argument supporting each P_j. Furthermore, by R we are warranted in believing that some

ticket will be drawn, so these conclusions conflict with one another. Intuitively, there is no reason to prefer some of the P_i's over others, so we cannot be warranted in believing any of them unless we are warranted in believing all of them. But we cannot be warranted in believing all of them, because the set $\{R,P_1,\ldots,P_{1000000}\}$ is inconsistent. In fact, it is a minimal inconsistent set. Hence by (4.7), we are not warranted in believing any of the P_i's.

4.4 Two Paradoxes of Defeasible Reasoning

The simple account of warrant that I gave in section 4.2 has some unacceptable consequences that will force its modification. This can be illustrated by looking at two apparent paradoxes of reasoning. First, let us look again at the lottery paradox. The lottery paradox is generated by supposing that we are warranted in believing a proposition R describing the lottery (it is a fair lottery, has one million tickets, and so on). Given that R is warranted, we get collective defeat for the proposition that any given ticket will not be drawn. But the present account makes it problematic how R can ever be warranted. Normally, we will believe R on the basis of being told that it is true (orally or in writing). In such a case, our evidence for R is statistical, proceeding in accordance with the statistical syllogism (3.10). That is, we know inductively that most things we are told that fall within a certain broad range are true, and that gives us a prima facie reason for believing R. So, we have only a defeasible reason for believing R. Let σ be the argument supporting R. Let T_i be the proposition that ticket i will be drawn. In accordance with the standard reasoning involved in the lottery paradox, we can extend σ to generate a longer argument η supporting $\sim R$. This is diagrammed in Figure 2. The final step of the argument is justified by the observation that if none of the tickets is drawn then the lottery is not fair.

The difficulty is now that η defeats σ by (4.1). Thus σ and η defeat one another, with the result that neither is ultimately undefeated. In other words, R and $\sim R$ are subject to collective defeat. This result is intuitively wrong. It should be possible for us to become warranted in believing R on the basis described.

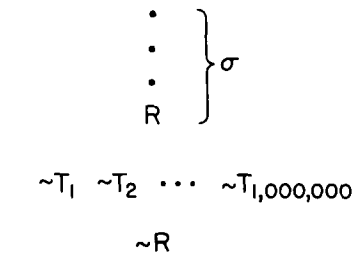

Figure 2.

Consider a second instance of paradoxical reasoning. Suppose we observe n A's r of which are B's, and then by statistical induction (3.13) we infer that prob(Bx/Ax)≈r/n. Supose that r/n is high. Then if we observe a further set of k A's without knowing whether they are B's, we can infer by statistical syllogism (3.10) that each one is a B. This gives us n+k A's of which r+k are B's. By (3.13), this in turn gives us a reason for thinking that prob(Bx/Ax)≈(r+k)/(n+k). If k is large enough, (r+k)/(n+k)≉r/n, and so we can infer that prob(Bx/Ax)≠r/n, which contradicts our original conclusion and undermines all of the reasoning. Making this more precise, we have two nested arguments diagrammed in Figure 3. σ and η defeat one another by (4.1), so we have a case of collective defeat. But this is intuitively wrong. All we actually have in this case is a reason for believing that prob(Bx/Ax)≈r/n, and a bunch of A's regarding which we do not know whether they are B's. The latter should have no effect on our warrant for the former. But by (4.1), it does. I will call this *the paradox of statistical induction.*

I believe that these two paradoxes illustrate a single inadequacy in our analysis of warrant. In each case we begin with an argument σ supporting a conclusion P, and then we extend σ to obtain an argument η supporting ~P. By (4.1), this is a case of collective defeat, but intuitively it seems that P should be warranted. What I think is happening here is that argument η is faulty all by itself. It is *self-defeating,* and that removes it from contention in any contest with conflicting arguments. Thus, it cannot enter into collective defeat with other arguments, and in particular it cannot enter into collective defeat with σ.

Let us define more precisely:

(4.5) σ is *self-defeating* if and only if σ supports a defeater for one of its own defeasible steps, that is, for some i and j, $<\sigma,i>$ defeats $<\sigma,j>$.

In order to handle the paradoxes, it suffices to remove self-defeating arguments from competition with other arguments. This can be done by revising

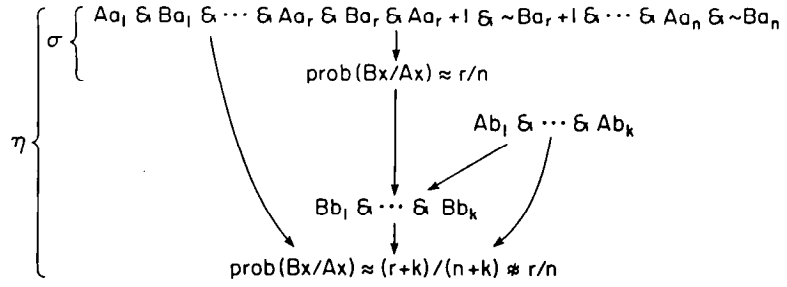

Figure 3.

the definition of "level 0 argument" to require that level 0 arguments be non-self-defeating. With this modification to the definition of "level 0 argument," I believe that our theory of warrant for linear arguments is correct. The principle (4.4) of collective defeat then comes out true as formulated, but the paradoxes of defeasible reasoning are no longer cases of collective defeat because the paradoxical arguments are no longer level 0 arguments.

5. A GENERAL THEORY OF WARRANT

5.1 A General Analysis

Now let us abandon the simplifying assumption that all arguments are linear. Indirect arguments differ from linear arguments in that they can adopt unestablished premises as suppositions and then later discharge those premises. This gives them a more complicated form than linear arguments. To accommodate this we must generalize the form of arguments. For this purpose, let us take a line of an argument to be an ordered quadruple $<\Gamma,P,R,\{m,n,\ldots\}>$ where Γ is a finite set of propositions (the *premise set* of the line), and P, R, and $\{m,n,\ldots\}$ are as before. $p(\sigma,i)$ is the premise set of the ith line of σ. We say that an argument *supports* a conclusion P if and only if P is supported by some line of the argument having an empty premise set.

In constructing a theory of warrant for nonlinear arguments, we must decide how subsidiary arguments are to be integrated into their enclosing arguments. In actual reasoning, we often import conclusions supported by the enclosing argument into the subsidiary arguments without defending them anew. This is because to defend the conclusion anew within the subsidiary argument would just involve repeating the original argument intact. However, there must be constraints on the extent to which we can do this. Suppose, for example, that the subsidiary argument proceeds from the supposition P, but the enclosing argument supports $\sim P$. Then obviously we cannot simply import $\sim P$ into the subsidiary argument without making that argument self-defeating. More generally, if Q is supported by the enclosing argument, we cannot import Q into the subsidiary argument if the subsidiary argument supports a defeater for some step that is involved in getting Q. If we attempted to formulate a constraint to handle this, it would be complicated and it would not be obvious at this point whether we had it right. This suggests that a better alternative for the theory of warrant is to preclude ever importing conclusions from the enclosing argument into the subsidiary argument. Instead, we require them to be established anew within the subsidiary argument, even if that involves simply repeating the earlier argument intact. We lose nothing in terms of supportability by this strategy, and we avoid the serious danger of getting the rules wrong by trying to make them too elegant. Remember that the theory of warrant is not intended to

be a theory of reasoning. Given this more conservative characterization of warrant, it becomes a substantive question with an objectively determinable answer just when conclusions supported by the enclosing argument can be reproduced within the subsidiary argument without rendering the whole argument self-defeating. The answer to this question can then be used in constructing more elegant rules for reasoning.

For nonlinear arguments we must modify our previous rules of inference and augment them with new rules. Given our conservative approach to the embedding of subsidiary arguments, rule **F** and rule **R** can be rewritten as follows:

Rule **F**: *Foundations*
If P expresses a foundation state contained in the espitemic basis, and Γ is any finite set of propositions, $<\Gamma,P,\mathbf{F},\emptyset>$ can be entered as any line of the argument.

Rule **R**: *Closure under reasons*
If $\{P_1,\ldots,P_n\}$ is a reason for Q and $<\Gamma,P_1,\ldots>,\ldots,<\Gamma,P_n,\ldots>$ occur as lines $i_1,\ldots,i_n$ of an argument, $<\Gamma,Q,\mathbf{R},\{i_1,\ldots,i_n\}>$ can be entered on any later line.

In addition we will have at least the following two rules governing indirect arguments:

Rule **P**: *Premise introduction*
For any finite set Γ and any P in Γ, $<\Gamma,P,\mathbf{P},\emptyset>$ can be entered as any line of the argument.

Rule **C**: *Conditionalization*
If $<\Gamma\cup\{P\},Q,\ldots>$ occurs as the *i*th line of an argument then $<\Gamma,(P\supset Q),\mathbf{C},\{i\}>$ can be entered on any later line.

Conditionalization is a very pervasive form of inference. There are a number of different kinds of conditionals—material, indicative, subjunctive, and so forth—and what is perhaps characteristics of conditionals is that they can all be inferred by some form of conditionalization. It can be shown that any conditional satisfying both rule **C** and modus ponens is equivalent to the material conditional,[8] but many kinds of conditionals satisfy weaker forms of conditionalization. In particular, I will assume the following weak form of conditionalization, related to the conditionals involved in undercutting defeaters:

---

[8] The proof is simple. Given such a conditional "$\rightarrow$," suppose $\{P,(P\supset Q)\}$. By modus ponens for "$\supset$," we get Q, and then by strong conditionalization for "$\rightarrow$," "$(P\rightarrow Q)$" follows from the supposition $\{(P\supset Q)\}$. A second conditionalization (this time with respect to "$\supset$" gives us $[(P\supset Q)\supset(P\rightarrow Q)]$." Conversely, using modus ponens for "$\rightarrow$" and strong conditionalization for "$\supset$," we get "$[(P\rightarrow Q)\supset(P\supset Q)]$." So "$[(P\supset Q)\equiv(P\rightarrow Q)]$"becomes a theorem.

Rule **WC**: *Weak conditionalization*
If $<\{P\},Q,\ldots>$ occurs as the ith line of an argument then $<\emptyset,(P\rightarrow Q)$, **SC**, $\{i\}>$ can be entered on any later line.

The difference between conditionalization and weak conditionalization is that the latter requires you to discharge all your assumptions at once. Given modus ponens and the principle of exportation:

$$[(P \,\&\, Q)\rightarrow R]\supset[P\rightarrow(Q\rightarrow R)]$$

conditionalization and weak conditionalization would be equivalent, but no conditional other than the material conditional seems to satisfy exportation.

The preceding does not comprise a complete list of the rules of inference used in human argumentation. At the very least it must be augmented with some rule for quantifiers. But this partial list will be sufficient for present purposes.

Our conservative approach to the construction of indirect arguments precludes importing conclusions from the enclosing arguments into subsidiary arguments. This makes it easy to give a characterization of defeat for indirect arguments. A line of one argument can defeat a line of another argument only if they have the same premise sets:

(5.1) $<\eta,j>$ *defeats* $<\sigma,i>$ if and only if (1) $<\sigma,i>$ is obtained by rule **R** using $\{P_1,\ldots,P_n\}$ as a prima facie reason for Q, (2) η_j is either $\sim Q$ or $\sim[(P_1 \,\&\ldots\&\, P_n)\rightarrow Q]$, and (3) $\mathbf{p}(\eta,j)=\mathbf{p}(\sigma,i)$.

Warrant is then characterized just as before in terms of successive levels of defeat and reinstatement. As before, a self-defeating argument is one in which one line defeats another, and we count every non-self-defeating argument as a level 0 argument. Then we define the notions of a level n argument and an ultimately undefeated argument just as before, and a proposition is warranted if and only if it is supported by an ultimately undefeated argument.

5.2 Logical Properties of Warrant

Our analysis of warrant enables us to prove that the set of warranted propositions has a number of important properties. Let us symbolize "P is warranted relative to the epistemic basis E" as "$\models\!\!>_E P$." Let us also define "warranted consequence":

(5.2) $\Gamma \models\!\!>_E P$ if and only if there is an ultimately undefeated argument relative to E that contains a line of the form $<\Gamma,P,\ldots>$.

We can prove that these notions have a number of important logical properties. Let us say that a proposition P is a *deductive consequence* of a set Γ of propositions (symbolized '$\Gamma \vdash P$') if and only if there exists a deductive argument leading from members of Γ to the conclusion P. I have assumed

that there are enough conclusive reasons to allow us to carry out deductive reasoning in terms of them. This has the consequence:

(5.3) If $\Gamma \vdash P$ then $\Gamma \models\!\!>_E P$.

I will say that a set of propositions is *deductively consistent* if and only if it does not have an explicit contradiction as a deductive consequence. The set of warranted propositions must be deductively consistent. (I assume here and throughout that an epistemic basis must be consistent.) If a contradiction could be derived from it, then reasoning from some warranted propositions would lead to the denial (and hence defeat) of other warranted propositions, in which case they would not be warranted. More generally:

(5.4) If Γ is deductively consistent so is $\{P|\ \Gamma \models\!\!>_E P\}$.

The set of warranted propositions must also be closed under deductive consequence:

(5.5) If for every P in Γ, $\models\!\!>_E P$, and $\Gamma \vdash Q$, then $\models\!\!>_E Q$.

To see this, suppose $P_1,\ldots,P_n$ are warranted and Q is a deductive consequence of them. Then an argument supporting Q can be constructed by combining arguments for $P_1,\ldots,P_n$ and adding onto the end an argument deducing Q from $P_1,\ldots,P_n$. The last part of the argument consists only of deductive nondefeasible steps of reasoning. If Q is not warranted, there must be an argument defeating the argument supporting Q. There can be no defeaters for the final steps, which are nondefeasible, so such a defeater would have to be a defeater for an earlier step. But the earlier steps all occur in the arguments supporting $P_1,\ldots,P_n$, so one of those arguments would have to be defeated, which contradicts the assumption that $P_1,\ldots,P_n$ are warranted. Thus, there can be no such defeater, and hence Q is warranted.

More generally:

(5.6) If for every P in Γ, $\Lambda \models\!\!>_E P$, and $\Gamma \models\!\!>_E Q$, then $\Lambda \models\!\!>_E Q$.

We also have the following analogue of the standard deduction theorem in classical logic:

(5.7) If $\Gamma \cup \{P\} \models\!\!>_E Q$ then $\Gamma \models\!\!>_E (P \supset Q)$.

This follows more or less immediately from rule **C**, and contrasts with the nonmonotonic logic of McDermott and Doyle (1980).

5.3 The Principle of Collective Defeat

The principle (4.4) of collective defeat remains true in our general theory of warrant, and its proof remains essentially unchanged. At this point, I want to observe that it has an interesting analogue. Principle (4.4) is a principle of *collective rebutting defeat.* It only pertains to cases in which we have argu-

ments supporting both P and ~P. But we can obtain a *principle of collective undercutting defeat* in precisely the same way:

(5.8) If we are warranted in believing R and there is a set Γ of propositions such that:
(1) we have equally good defeasible arguments for believing each member of Γ;
(2) for each P in Γ, the supporting argument involves a defeasible step proceeding from some premises $S_1,\ldots,S_n$ to a conclusion T, and there is a finite subset Γ_P of Γ such that the conjunction of R with the members of Γ_P provides a deductive argument for $\sim[(S_1,\ldots,S_n)\rightarrow T]$ that is as strong as our initial argument is for P; and
(3) none of these arguments is defeated except possibly by their interactions with one another;
then none of the propositions in Γ is warranted on the basis of these defeasible arguments.

A simple illustration of this principle will involve a pair of arguments having the structure diagrammed in Figure 4. For instance, R might be "People generally tell the truth." Suppose P is "Jones says Smith is unreliable" and Q is "Smith says Jones is unreliable." By statistical syllogism (3.10), (P & R) is a prima facie reason for believing S: "Smith is unreliable"; and (Q & R) is a prima facie reason for believing T: "Jones is unreliable." But S is an undercutting defeater for the reasoning from (Q & R) to T, and T is an undercutting defeater for the reasoning from (P & R) to S. Presented with this situation, what should we believe about Smith and Jones? The intuitive answer is, "Nothing." We have no basis for deciding that one rather than the other is unreliable. Under the circumstances, we should withhold belief regarding their reliability. And that is just what principle (5.8) tells us.

We have separate principles of collective defeat for rebutting defeaters and undercutting defeaters. Can we also have mixed cases of collective defeat involving both rebutting and undercutting defeaters? Interestingly, that does not seem to be possible. Such cases would have the form diagrammed in Figure 5, or a generalization of that form, where all steps are defeasible. The idea here is that σ provides rebutting defeat for η, and η provides under-

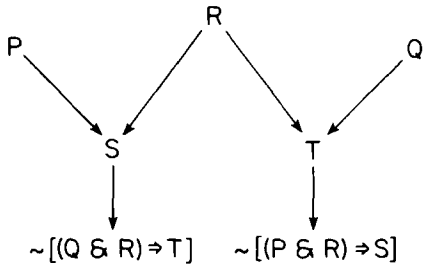

Figure 4

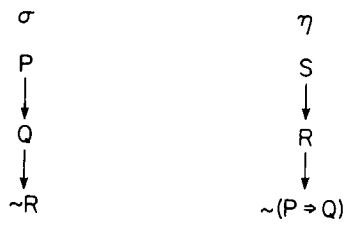

Figure 5.

cutting defeat for σ. But the relation "provides rebutting defeat for" is symmetrical, so η also provides rebutting defeat for σ. Thus, this is a simple case of collective rebutting defeat, and the last step of η is irrelevant.

5.4 The Principle of Joint Defeat

Sometimes we are in a position of having a disjunction of defeaters warranted without the individual defeaters being warranted. The principle of joint defeat tells us that in such a case, the defeaters defeat even without being warranted. We have the following theorem:

(5.9) Suppose P is a prima facie reason for Q and R is a prima facie reason for S, and T = "$\sim$(P$\rightarrow$Q)" and U = "$\sim$(R$\rightarrow$S)". If "T v U" is warranted but neither T nor U is warranted, then no argument using either P as a reason for Q or R as a reason for S (in accordance with rule **R**) is ultimately undefeated.

Proof: Suppose σ is an argument using one of these reasons. For specificity, suppose σ uses R as a reason for S. Suppose μ is an ultimately undefeated argument for "T v U." We can construct an argument η by adding the following lines to the end of μ:

(i) $<\emptyset, \sim(P\rightarrow Q)\ v\ \sim(R\rightarrow S), \ldots>$ (last line of μ)
(i+1) $<\{P\},P,\mathbf{P},\{i+1\}>$
(i+2) $<\{P\},Q,\mathbf{R},\{i+1\}>$
(i+3) $<\emptyset,P\rightarrow Q,\mathbf{WC},\{+2\}>$
(i+4) $<\emptyset, \sim(R\rightarrow S)$, deductive inference, $\{i,i+3\}>$

Similarly, we can construct an argument γ for $<\phi, \sim(P\rightarrow Q), \ldots>$. η and γ defeat one another, so each fails to be a level n argument for odd n, but each is a level n argument for all even n. η defeats σ, so σ also fails to be a level n argument for any odd n. Thus σ is not ultimately undefeated.

More generally, we have:

(5.10) Given a finite set of triples $<P_i,Q_i,R_i>$ where P_i is a prima facie reason for Q_i and R_i = "$\sim(P_i\rightarrow Q_i)$," if the disjunction of all the R_i's is warranted but the disjunction of every proper subset of the R_i's is unwarranted, then no argument using P_i as a reason for Q_i in an application of rule **R** is ultimately undefeated.

This is *the principle of joint defeat*. The principle can be generalized in various ways. For example, it can be combined with collective undercutting defeat to give us cases of collective joint defeat.

5.5 Prima Facie Warrant and Default Assumptions

AI studies of nonmonotonic reasoning have focused on reasoning from default assumptions. By contrast, the nonmontonicity of defeasible reasoning results from the operation of prima facie reasons. The latter can be viewed as a kind of default assumption, but they are not assumptions in the sense of beliefs. For instance, in making a perceptual judgment about an object X, human beings do not usually have the explicit belief that if X looks red then it is red. The prima facie reason plays a *processing* role. It is a constituent of the production system governing reasoning, and it can play that role without the corresponding conditional having any explicit representation in thought.

Default assumptions in familiar AI programs play a role analogous to beliefs that are held until they must be given up. We can define a similar notion in the theory of warrant:

(5.11) P is *prima facie warranted* if and only if, relative to every epistemic basis, P is automatically warranted in the absence of a reason for believing $\sim P$.

Prima facie warranted propositions are much like the traditional construal of default assumptions, except that they can be prima facie warranted without being believed.

Are there any prima facie warranted propositions? It is easily established that there are. If P is a prima facie reason for Q, then by **WC** we are automatically warranted in believing $(P \rightarrow Q)$ unless we have an undercutting defeater for the prima facie reason. But an undercutting defeater is just a reason for denying the conditional, so it follows that the conditional is prima facie warranted.

Are there any other prima facie warranted propositions? At one time, I thought that the propositions comprising the epistemic basis were prima facie warranted (Pollock, 1974), but I have recently argued that they are not (Pollock, 1986). I suspect that the only legitimate examples of prima facie warrant are conditionals derived from prima facie reasons. Still, the notion may prove useful for AI for the practical purpose of modeling everyday reasoning. Ordinary default assumptions are not really prima facie warranted propositions. Rather, they are beliefs based upon defeasible reasons. But for practical purposes we may not want to trace the epistemological connections all the way back back to the epistemic basis. Thus, it may be useful to model much everyday reasoning by pretending that various "high level" beliefs are prima facie warranted.

Although some of the applications of default reasoning in AI can be modeled on defeasible reasoning in the manner I have suggested, much cannot. The use of default assumptions in AI is often aimed at planning (making airline reservations, scheduling meetings, etc.) rather than believing. We can plan to do something in a certain way *if we can* (i.e., schedule the meeting on Wednesday if possible). It does not seem that this can be subsumed under defeasible reasoning as a theory of belief formation. The use of defaults in planning is actually the more general of the two and subsumes defeasible reasoning. Defeasible reasoning can be regarded as proceeding in terms of default instructions of the form, "Reason this way if you can." The very fact of this subsumption, combined with the complex logical structure of defeasible reasoning, seems to indicate the default reasoning in planning must, in general, have a more complex structure than has been recognized heretofore.

The fact that the conditionals corresponding to prima facie reasons are prima facie warranted suggests an alternative reconstruction of human reasoning. Rather than taking the prima facie reasons to be the basic constituents of reasoning, why not take the prima facie warranted conditionals to be basic and describe the reasoning as proceeding deductively (by modus ponens) from the conditionals? Then the only thing that would have to be defeasible about the reasoning would be the warrant of the premises. As a purely psychological observation, human reasoning does not proceed in this way. As I remarked above, humans do not explicitly store the conditionals as beliefs. Reasons play the role of rules of inference rather than premises. This may seem like a rather incidental feature of human reasoning—a "psychological accident" so to speak. It may seem that we could build an intelligent machine that reasoned in the manner described rather than in the human fashion, and the resulting reasoning would look much more like standard AI accounts of nonmonotonic reasoning. This, however, is false. The difficulty is that reasons come in schemas rather than as individual reasons, and the schemas encompass infinitely many cases. For instance, "X looks red to me" is a prima facie reason for me to believe "X is red," and this is true regardless of what term X is. Replacing this reason schema by explicitly stored prima facie warranted conditionals would require storing infinitely many conditionals, and that is impossible. It might seem that we could instead take the universal generalization "$(\forall x)$[(X looks red to me)$\rightarrow$(X is red)]" to be prima facie warranted, and store it. But this would not do the same job. The conditional would be defeated by finding a single case of an object that looks red but is not red, and once that happens the conditional would not warrant any further inferences regarding the colors of other objects. In contrast to this, different instances of prima facie reasoning regarding the colors of different objects are relatively independent of one another. Defeating one does not defeat the others, at least until we get so

many counter instances that we are warranted in concluding inductively that this prima facie reason is generally unreliable. So it seems the defeasible reasoning cannot be replaced by default reasoning from prima facie warranted propositions.

PART III

6. JUSTIFIED BELIEF AND RULES FOR REASONING

I glossed "warrant" as "what an ideal reasoner would believe." An ideal reasoner is one unconstrained by a finite memory or processing capacity. Warrant is an ideal to which "real" epistemic agents aspire. But we cannot expect real epistemic agents to believe only warranted propositions. Warrant is a "global" concept defined in terms of the set of all possible arguments available to an epistemic agent at a single time. No one can actually survey that infinite totality and decide what to believe by applying the definition of "warrant" to it. That definition involves the comparison of infinitely many arguments and, in cases of collective defeat, the infinite cycling of arguments through defeat and reinstatement. This could not reflect the way we actually reason. Actual rules for reasoning must appeal exclusively to "local" considerations—readily accessible features of the epistemic situation.

Insofar as we reason in accordance with our built-in rules for reasoning, whatever they may be, our beliefs are said to be *justified,* but this does not guarantee that they are warranted. Justification only approximates warrant. We can, for example, be justified in holding deductively inconsistent beliefs if we are unaware that they are inconsistent and we got to them in reasonable ways, but deductively inconsist beliefs can never be warranted. Warrant is at most an ideal to which justified reasoning aspires. Actual reasoning takes the form of working out arguments and defeating and reinstating them in the same manner as is involved in the definition of warrant, but we are limited in how many arguments and how many steps of defeat and reinstatement we can go through. If we could keep going indefinitely, our rules for reasoning would lead to warranted beliefs, but of course, we cannot. Instead, our rules for justified belief formation involve the presumption that the reasoning we have done at any given time is "all right" unless we have some concrete reason, in the form of a defeating argument, for thinking otherwise. This is a kind of second-order default assumption about the existence of defeating arguments.

Let us try to be more specific about the rules for justified belief formation. In this section I will make some general remarks about the form of the rules for defeasible reasoning, and in the next section I will attempt to construct concrete rules of this form.

The most natural assumption is that rules for reasoning tell us to form certain beliefs if we already have other appropriately related beliefs. Such

rules *prescribe* the formation of beliefs whenever we believe the premises from which they can be obtained. But actual rules for reasoning do not take this form. The simplest rules are "permission rules," telling us that it is *all right* to form various kinds of beliefs, but not that we *must* form them. For example, a common mistake is to formulate a rule of *modus ponens* as follows:

(6.1) If one believes both P and $(P \supset Q)$ then he should believe Q.

But this is wrong. One does not automatically have an epistemic obligation to believe everything he can infer from his various beliefs. As Gilbert Harman (1986) has observed, such an obligation would lead to unmanageable clutter in one's beliefs. Normally, one only draws conclusions insofar as one is interested in the conclusions or there is reason to believe that the conclusions bear upon matters that interest one. Thus *modus ponens* should be at most a permission rule, *allowing* us to draw a certain conclusion rather than mandating our drawing that conclusion.[9] It is such epistemic permission rules that have been the focus of much work in epistemology.

On the other hand, a production system for belief formation must tell us to adopt particular beliefs under various circumstances. It cannot just tell us to do so if we want. Thus, there must be rules prescribing belief formation, but what the above observations indicate is that these rules must appeal not just to what other beliefs we have, but also to what our interests are. This suggests that rules prescribing belief formation must have forms more like the following:

(6.2) If you have beliefs $P_1, \ldots, P_n$ and you are interested in whether Q is true then you should believe Q.

For now, I will not worry about how one becomes interested in various propositions, although this is a matter that must eventually be addressed.

It appears that there must be three kinds of rules for belief formation. First, there are *adoption rules* telling us that we should adopt beliefs if we care about whether they are true and we have arguments supporting them. These rules create a "prima facie obligation" to adopt beliefs. Second, there are *defeat rules* canceling that prima facie obligation when we discover defeating arguments for the initial arguments. This second category of rules consists of obligation rules prescribing the withholding of belief. Again, these rules create only prima facie obligations, because defeating arguments can themselves be defeated. This leads to the third class of rules—the *reinstatement rules*—which concern reinstatement from defeat. Reinstatement occurs when defeaters are themselves defeated.

---

[9] This is still simplistic, because sometimes what we should do is reject either P or $(P \supset Q)$ rather than coming to believe Q.

There are two kinds of candidates for adoption rules. The simplest proposal would be that whenever $\{P_1,\ldots,P_n\}$ is a reason for Q, we have a rule of roughly the form:

(6.3) If you believe $P_1,\ldots,P_n$ and you have no defeating beliefs and you care whether Q then you should believe Q.

Alternatively, observing that when you believe Q on the basis of $P_1,\ldots,P_n$, your belief in Q is not justified unless your belief in $P_1,\ldots,P_n$ is justified, it might be insisted that correct rules should have the form:

(6.4) If you justifiably believe $P_1,\ldots,P_n$ and you have no defeating beliefs and you care whether Q then you should believe Q.

However, it seems that in order for a system to implement a rule of the form of (6.4) it would first have to form beliefs about how it came to believe $P_1,\ldots,P_n$. This would lead to an infinite regress. To avoid any such regress, rules for belief formation must be local in the sense that they can be instantiated without first forming other beliefs. A system can instantiate (6.3) without first forming the belief that it believes $P_1,\ldots,P_n$, because a system can be built to respond directly and "nondoxastically" to what beliefs it has. But without a computationally impractical amount of recordkeeping it could not similarly respond directly to the justificational status of its beliefs. Human beings do not keep track of their arguments to any great extent. The production system for human reasoning just assumes that beliefs are justified until proven otherwise. Furthermore, it seems that a reasoning machine must work similarly. If it had to keep track of all its arguments its memory would rapidly become so cluttered that memory searches would be immense tasks and would slow reasoning down to a crawl.[10] So in other words, our positive rules for belief formation take the form of (6.3).

Our defeat rules pertain to the discovery of defeaters for the arguments on the basis of which beliefs are held. A natural hypothesis is that whenever $\{P_1,\ldots,P_n\}$ is a prima facie reason for Q and R is a reason for either $\sim Q$ or $\sim[(P_1 \&\ldots\& P_n)\rightarrow Q]$, we must have a rule something like the following:

(6.5) If you believe Q on the basis of an argument one line of which is obtained by rule **R** using $\{P_1,\ldots,P_n\}$ as a prima facie reason for Q, then if you adopt R as a new belief, you should cease to believe Q on this basis.

But the implementation of this rule requires the system to keep track not only of its beliefs but also of the arguments on the basis of which it holds the beliefs, and we have seen that that is impractical. This suggests that in place of (6.5), what we actually have is a "doxastic" rule more like:

---

[10] In this connection, compare Doyle's (1979) truth maintainance system, which does keep track of all its arguments.

(6.6) If (1) you believe that you believe Q on the basis of an argument one line of which is obtained by using P as a prima facie reason for Q, and (2) you believe a defeater, then you should cease to believe Q on this basis.

I am, however, uncomfortable with (6.6). (6.6) has the consequence that defeat requires higher-order monitoring of our reasoning processes. It is indisputable that such higher-order monitoring sometimes occurs, but it is a complicated process and I do not think that defeat requires it. Defeat often proceeds in a more automatic fashion. If I believe Q on the basis of a prima facie reason P, and then I adopt a new belief R that is a prima facie reason for $\sim$Q, I just automatically retract Q without having to think about the matter. How to construct rules for defeat that work in such an automatic fashion without higher-order monitoring is one of the problems that I will face in the next section. In a recent book (Pollock, 1986) I argued that this was impossible, but it turns out to be fairly easy.

A major problem that must be faced in the construction of rules for reasoning is that they must avoid infinite cycling. The theory of warrant handles collective defeat in terms of infinite cycling between competing arguments, but actual reasoning must work in some simpler fashion.

With these preliminary remarks as a background, I turn now to the construction of a concrete system of rules for defeasible reasoning. This system, and the computer program implementing it, will be called "OSCAR."[11]

7. OSCAR: A FRAMEWORK FOR DEFEASIBLE REASONING

The problem of constructing a general framework for defeasible reasoning is a difficult one, and I do not at this time have an entirely general solution to the problem. Instead, my strategy is to adopt some simplifying assumptions and construct a theory of defeasible reasoning based on those assumptions. My strategy for future research will then be to remove the simplifying assumptions one at a time, each time making the theory more sophisticated in order to handle the greater generality.

I will make the following simplifying assumptions:

1. The most important simplification will result from confining my attention to linear arguments. I will pretend that all arguments are linear so that we do not have to contend, for example, with conditional arguments.
2. I will take the system to be interested in everything, so that it draws all possible conclusions. In order for this to work, we must be careful what reason schemes we make available to OSCAR. If we give him all of

[11] 'Oscar' is the hero in "The Fable of Oscar" (Chapter Five of Pollock, 1986). For the further significance of OSCAR, see my (in press).

logic, the result will be a combinatorial explosion. He could, for example, spend all his time forming longer and longer conjunctions out of just two initial beliefs. The general framework of OSCAR will impose no constraints on reasons, but in actually running OSCAR we will have to observe constraints. There is no cure for this short of building in interest constraints, and that must await further developments in OSCAR.

3. My final simplifying assumption will be that all reasons have the same strength. This has the consequence that whenever we have a reason for P and a reason for $\sim$P, we have a case of collective defeat. Of course, this consequence is not realistic, because in actual practice we can have a strong reason for P and a weak reason for $\sim$P, and still be at least mildly justified in believing P.

I will now discuss a system of rules for defeasible reasoning based upon these assumptions, and then I will discuss briefly how the rules will have to be modified in order to relax these assumptions.

Any realistic system of reasoning must be "computationally feasible," in the sense that it prescribes reasoning that can be carried out in realistic amounts of time. In order to accomplish this, it must avoid various sources of "combinatorial explosion." For example, Gilbert Harman's (1984) well known objection to basing reasoning on condtional probability amounts to the observation that it requires storing and retrieving unreasonably large sets of probabilities. Some of the reasoning systems that have been proposed in AI are subject to the same difficulty. For example, Jon Doyle's *truth maintenance system* (1979) requires the reasoner to keep track of all of his arguments and be able to access them for purposes of defeat and reinstatement. But that puts a tremendous burden on memory, because storing an argument takes much more memory than storing a belief, and as beliefs build upon one another in a cumulative fashion their supporting arguments become progressively longer. Human beings accomplish defeasible reasoning without being very good at remembering arguments.

In the interest of computational efficiency, it is desirable for a system to avoid, insofar as possible, having to search its full set of beliefs. This can be done by storing newly adopted beliefs separately from other beliefs, and searching only the set of newly adopted beliefs in determining what new reasoning it should carry out. In this connection, OSCAR assumes that anything that can be inferred from old beliefs has already been inferred. Thus, all beliefs are stored in the set *beliefs,* and newly adopted beliefs are stored in the set *adoptionset.* Each time a cycle of reasoning is completed, *adoptionset* is cleared. The rules for belief adoption will then have the form:

> Where p is a reason for q, if p$\in$*adoptionset* and this reason is undefeated, then adopt q.

The major problem that I have addressed in this version of OSCAR is how to handle defeat and reinstatement without storing all of the arguments

on the basis of which beliefs are adopted. It turns out that this can be done by storing only the "immediate bases" for beliefs, and the bases for defeat. More specifically, rather than storing the entire argument, we just store the last step. If we inferred Q from P, we store that fact, but no more of the argument. Of course, human beings do sometimes remember their arguments, but my concern here is to construct a "basic" system of defeasible reasoning that can accomplish its goals with as little higher-order monitoring as possible. Given such a system, we can consider later how it might be streamlined by allowing higher-order monitoring to play a role when it is available.

In addition to keeping track of the reason for holding a belief, we must keep track of whether that reason is defeasible or conclusive. This makes a difference to the way in which defeat works (see particularly the rule BACK-TRACK below). Taking explicit account of the fact that reasons are sets of propositions, let us define:

(7.1) S *believes* P *on* X if and only if S believes P on the basis of the defeasible reason X.

(7.2) S *believes* P *con* X if and only if S believes P on the basis of the conclusive reason X.

These are to be understood in such a way that an agent can believe something on or con several different bases at the same time. I will take *onset* to be the set of pairs $<Q,X>$ such that Q is believed on X, and *conset* will be the set of pairs $<Q,X>$ such that Q is believed con X.

We must also keep track of the bases upon which reasons are defeated. In the case of undercutting defeat, that is simple. Suppose X (a set of propositions) is a prima facie reason for Q, and $\wedge X$ is the conjunction of all the members of X. If this reason is defeated by undercutting, then the agent believes $\sim(\wedge X \rightarrow Q)$, and the basis for that is stored in *onset* or *conset*. Reinstatement then results from anything leading the agent to retract belief in $\sim(\wedge X \rightarrow Q)$.

Rebutting defeat is more complicated. Consider a pair of arguments leading to contradictory conclusions, as in Figure 6, where {P} is a prima facie reason for R, {Q} is a prima facie reason for S, and the remaining reasons are conclusive. As this is a case of collective defeat, neither V nor $\sim V$ should be believed. When we encounter collective defeat, we must do more than just reject the conclusions. It must be possible for a collectively defeated argument to be reinstated by its competitors becoming defeated. Thus, for instance, if we acquire an undercutting defeater for the move from P to R, that should reinstate the argument from Q to $\sim V$. However, given that we only search newly adopted beliefs in deciding what new beliefs to adopt, we will not automatically "rediscover" the argument from Q to $\sim V$ after defeating the argument from P to V. Thus, we must store facts about collective defeat in a way that will make reinstatement possible. It will not suffice to

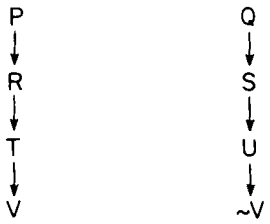

Figure 6.

just store the sets of propositions that enter into collective defeat, because we must also know what it takes to defeat some and reinstate others. A proposition is always defeated by defeating the *last defeasible step* of the argument supporting it. The last defeasible step can be defeated either by retracting belief in the premise or by undercutting the prima facie reason involved. Thus when a set of arguments collectively defeat one another, what OSCAR will store is the set of their last defeasible steps. For instance, in the above example, OSCAR will store the set $\{<R,\{P\}>,<S,\{Q\}>\}$. Reinstatement is then accomplished by defeating one of those last defeasible steps, readopting the conclusions of the others, and then repeating the earlier purely deductive reasoning that initially led to collective defeat. Thus if OSCAR acquires a defeater for $<R,\{P\}>$, he will readopt belief in S, and that will automatically lead him to repeat the deductive reasoning leading to the conclusion $\sim V$. OSCAR will keep track of such rebutting defeat by putting the pair $\{<R,\{P\}>,<S,\{Q\}>\}$ in the set *rebut*.

In rebutting defeat, we must backtrack to the last defeasible step of reasoning and then take that to be rebutted. This is made precise in terms of the notion of a nearest defeasible ancestor. In the simple case where all reasons are unit sets, this notion can be defined as follows:

(7.3) Where p is a belief, q is a *nondefeasible ancestor* of p if and only if there is a sequence $<p_1,\ldots,p_n>$ $(i>1)$ of beliefs such that $p=p_n$, $q=p_1$, and for each $i<n$, $<p_{i+1},\{p_i\}>\epsilon$*conset*. A pair $<r,\{s\}>$ is a *nearest nondefeasible ancestor* of p if and only if r is a nondefeasible ancestor of p and $<r,\{s\}>\epsilon$*onset*.

In the general case, where reasons can be arbitrarily large finite sets, the definition is more complicated. First, define recursively:

(7.4) 1. If $<p,X>\epsilon$*conset* then X is a *nondefeasible ancestor-set* of p.
2. If x is a *nondefeasible ancestor-set* of p, $q\epsilon X$, and $<q,Y>\epsilon$*conset*, then $(X-\{q\})\cup Y$ is a *nondefeasible ancestor-set* of q.

(7.5) q is a *nondefeasible ancestor* of a proposition p if and only if q is a member of some nondefeasible ancestor-set of p.

(7.6) q is a *nondefeasible ancestor* of a set X if and only if q is a nondefeasible ancestor of some member of X.

(7.7) X is a *nearest defeasible ancestor* of a proposition p if and only if X is a set of ordered pairs, the domain of X is a nondefeasible ancestor-set of p, and $X \subseteq$ *onset*.

A nearest defeasible ancestor of p is (roughly) the set of bottom nodes of a deductive argument leading upward to p. It is these that are involved in collective defeat.

Finally:

(7.8) X is a *nearest defeasible ancestor* of $\{p_1,\ldots,p_n\}$ if and only if there are $X_1,\ldots,X_n$ such that for each i, X_i is a nearest defeasible ancestor for p_i, and $X = X_1 \cup \ldots \cup X_n$.

When two chains of reasoning lead to contradictory conclusions, we take the nearest nondefeasible ancestors to rebut one another, and we put the corresponding pairs of pairs in *rebut*. Let us define:

(7.9) An individual proposition P *is rebutted* if and only if for some set A in *rebut* and for some X, $<P,X> \epsilon A$.

I have already noted the OSCAR distinguishes between newly adopted beliefs and previously held beliefs. When he adopts a new belief, he then looks to see what consequences that has for his other beliefs. It may lead him to adopt further new beliefs, and it may lead him to retract old beliefs. OSCAR keeps track of newly adopted beliefs in *adoptionset*. Similarly, newly retracted propositions are put in the set *retractionset*. When a new belief is adopted, the rules of defeasible reasoning will be applied repeatedly in a certain order until no further adoptions or retractions can be obtained. This requires us to keep track of whether new adoptions or retractions occur at various points in the processing, and this will be done with the *adoption-flag* and the *retractionflag*, whose values are initially 0, but are reset to 1 whenever there is a new adoption or retraction. In deciding what to do at various points, the system will look at the value of these flags. If the flags are both 0 then no new adoptions or retractions have occurred. In that case, *adoptionset* and *retractionset* are cleared and the system is ready to process new inputs.

Let us define:

(7.10) *Adopting* a belief consists of (1) inserting it into *beliefs* if it is not already there, (2) putting it in *adoptionset*, (3) deleting it from *retractionset*, and (4) setting *adoptionflag* equal to 1. To *adopt P on X* is to insert $<P,X>$ in *onset* and adopt P. To *adopt P con X* is to insert $<P,X>$ in *conset* and adopt P.

Similarly:

(7.11) *retracting* a belief P consists of (1) deleting it from *beliefs*, (2) inserting it in *retractionset*, (3) deleting it from *adoptionset* if it is there, (4) deleting

all pairs of the form <P,X> from *onset*, and (5) setting *retractionflag* equal to 1. (For technical reasons connected with the rule (BACK-TRACK), we do not also delete pairs of the form <P,X> from *conset*.)

I remarked above that reasons come in schemas. For example, for any term X, "X looks red to me" is a prima facie reason for "X looks red." Here, "X looks red to me" and "X is red" express proposition-forms. In general, a reason schema is a pair <X,P> where P is a proposition-form and X is a set of proposition-forms.

Corresponding to a proposition-form p there are two functions whole_p, and parts_p. Applied to a proposition having the form p, parts_p generates an assignment of propositional constituents to the variables of p. Conversely, applied to such an assignment, whole_p generates the corresponding proposition. An assignment is a set of pairs <x,a> where X is a metalinguistic variable (a variable occurring in the formulation of proposition-forms) and a is the object assigned to that variable by the assignment. It is convenient to simply identify a proposition-form p with the pair of functions $<\text{whole}_p, \text{parts}_p>$. Given an assignment s, I will take p!s to be $\text{whole}_p(s)$, that is, the proposition of form p resulting from the assignment s. Similarly, where X is a set $\{p_1,\ldots,p_n\}$ of proposition-forms, I will take X!s to be $\{p_1!s,\ldots,p_n!s\}$.

Given these preliminaries, we now formulate our rules for adoption, defeat, and reinstatement precisely as follows:

Adoption

(ADOPT-ON)
Where <X,q> is a prima facie reason scheme:
For any variable assignment s, if you adopt X!s then if you do not believe $\sim(\wedge X!s \rightarrow q!s)$ and you do not believe $\sim q!s$ and neither q!s nor $\sim q!s$ is rebutted, then adopt q!s on X!s.[12]

(ADOPT-CON)
Where <X,q> is a conclusive reason scheme:
For any variable-assignment s, if you adopt X!s then if you do not believe $\sim q!s$ and neither q!s nor $\sim q!s$ is rebutted, then adopt q!s con X!s.

Retraction

(a) By undercutting defeat:

(UNDERCUT)
Where <X,q> is a prima facie reason scheme:
For any variable-assignment s, if you adopt $\sim(\wedge X!s \rightarrow q!s$ and you believe q!s on X!s, then delete <q!s,X!s> from *onset*, and retract q!s if you do not believe it on or con any other basis:

[12] In order to avoid complicating the rules throughout, I take $\sim\sim q$ to be q. That way we do not have to adopt replicas of our rules throughout for double negations.

(b) By rebutting defeat:

(REBUTa)
Where $<X,q>$ is a prima facie reason scheme:
For any variable-assignment s, if you believe $\sim$q!s and you adopt X!s, and either you do not believe $\sim(\wedge X!s \rightarrow q!s)$ or it is newly adopted, then find the nearest defeasible ancestors A of $\sim$q!s and retract $\sim$q!s and all the intermediate nondefeasible ancestors, and add the sets $A \cup \{<q!s,X!s>\}$ to *rebut*.

(REBUTb)
Where $<X,q>$ is a conclusive reason scheme:
For any variable-assignment s, if you believe $\sim$q!s and you adopt X!s then (1) find the nearest defeasible ancestors A of X!s and retract all members of X!s and all its intermediate nondefeasible ancestors, (2) find the nearest defeasible ancestors B of $\sim$q!s and retract $\sim$q!s and all its intermediate nondefeasible ancestors, and (3) add all the sets $A \cup B$ to *rebut*.

Enlarging Collective Defeat:

(NEG-COL-DEFa)
Where $<X,q>$ is a prima facie reason scheme:
For any variable-assignment s, if you adopt X!s and you do not believe $\sim(\wedge X!s \rightarrow q!s)$ and $\sim$q!s is rebutted, then for each Y such that $<\sim q!s, Y> \in$ *rebut*, add $\{<q!s,X!s>, <\sim q!s,Y>\}$ to *rebut*.

(NEG-COL-DEFb)
Where $<X,q>$ is a conclusive reason scheme:
For any variable-assignment s, if you adopt X!s and $\sim$q!s is rebutted, then find the nearest defeasible ancestors A of X!s and retract all the intermediate nondefeasible ancestors, and for each Y such that $<\sim q!s,Y>$ appears in *rebut*, add all the sets $A \cup \{<\sim q!s,Y>\}$ to *rebut*.

(POS-COL-DEFa)
Where $<X,q>$ is a prima facie reason scheme:
For any variable-assignment s, if you adopt X!s and do not believe $\sim(\wedge X!s \rightarrow q!s)$, and q!s is rebutted, then for each Y and A such that $<q!s,Y> \in A$ and $A \in$ *rebut*, add $\{<q!s,X!s>\} \cup (A - \{<q!s,Y>\})$ to *rebut*.

(POS-COL-DEFb)
Where $<X,q>$ is a conclusive reason scheme:
For any variable-assignment s, if you adopt X!s, and q!s is rebutted, then find the nearest defeasible ancestors A of X!s and retract all the intermediate nondefeasible ancestors, and for each B and Y such that $\{<q!s,Y>\} \in B$ and $B \in$ *rebut*, add the sets $A \cup (B - \{<q!s,Y>\})$ to *rebut*.

Hereditary Retraction:

(H-RETRACT)
For any X and q, if you believe q on X or q con X, but you retract some member of X, then retract q and delete $<q,X>$ from *conset* $\cup$ *onset*.

Backtracking for Conclusive Reasons:

(BACKTRACK)
For any X and q, if you believe q con X but you retract q, then retract X.

Reinstatement:
Reinstatement is handled by treating propositions as newly adopted (even if they have been believed all along) when sources of defeat are removed, and then seeing what can be inferred from them.

(a) from undercutting defeat:

(U-REINSTATE)
Where $<X,q>$ is a prima facie reason scheme:
For any variable-assignment s, if you believe X!s and retract $\sim(\wedge X!s \rightarrow q!s)$ and neither q!s nor $\sim q!s$ is rebutted, then adopt every member of X!s.

(b) from rebutting defeat:

by undercutting:

(R-REINSTATE/U)
Where $<X,q>$ is a prima facie reason scheme:
For any variable-assignment x, if you adopt $\sim(\wedge X!s \rightarrow q!s)$ and $<q!s, X!s> \epsilon A$ where $A \epsilon$ *rebut*, then delete A from *rebut* and adopt every member of $\cup$(range $(A - \{<q!s,X!s>\})$).

by retracting:

(R-REINSTATE/R)
If you retract some member of X and $\{<q,X>\} \epsilon A$ where $A \epsilon$ *rebut*, then delete A from *rebut* and adopt every member of $\cup$(range $(A - \{<q,X>\})$).

We can combine these rules straightforwardly in the architecture represented by Figure 7. There are more efficient architectures. This architecture involves needless repetitions of the searches, in the adoption module, for instances of reason schemes. However, these details are not important for the purposes of this paper.

8. ASSESSMENT OF OSCAR

How well does the present OSCAR perform? OSCAR is not an entirely realistic model of human defeasible reasoning, because it is based upon the simplifying assumptions listed at the beginning of the previous section. Nevertheless, OSCAR is a start at developing a realistic theory of defeasible reasoning. OSCAR does most things right, but suffers from some defects connected with those simplifying assumptions. One case worth mentioning

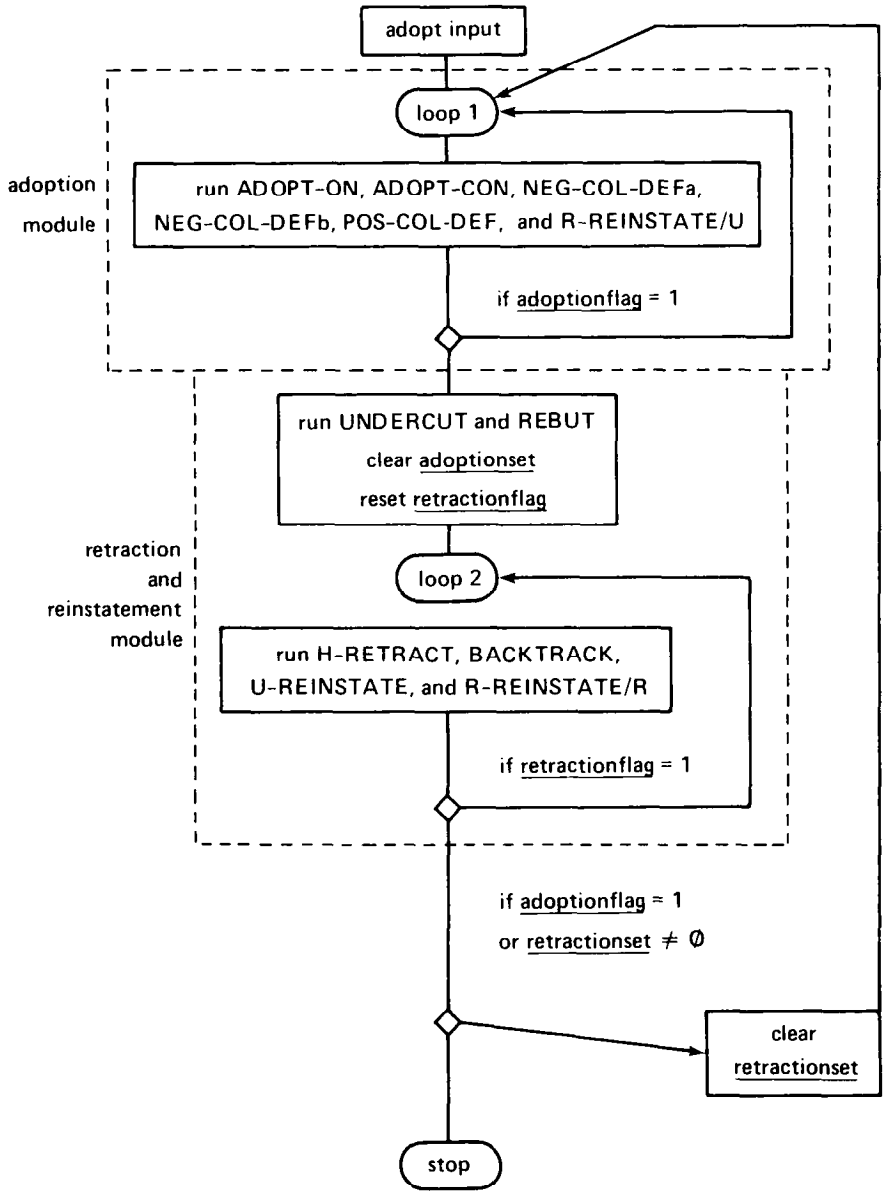

Figure 7.

is the following. Suppose A is a prima facie reason for P which is a conclusive reason for Q; B is a prima facie reason for R which is a conclusive reason for Q; and C is a prima facie reason for S which is a conclusive reason for $\sim$Q:

A- - ->P----->Q
B- - ->R----->Q
C- - ->S-----> ~Q

Given the input {A,C}, rebutting defeat occurs with the result that *beliefs* is just {A,C}, and *rebut* is [{<P,A>,<S,C>}]. So far so good, but now if we give OSCAR the new input {B}, we would like this to add B to *beliefs* (which it does) and add {<R,B>,<S,C>} to *rebut*. The latter fails. Instead, OSCAR adopts R on B and then adopts Q con R. This is because there is no mention of Q itself in *rebut*. Instead, it is the nearest defeasible ancestors of Q that went into *rebut*. There is an *ad hoc* way of handling this in OSCAR. If S is a conclusive reason for ~Q then S entails ~Q, and so Q entails ~S. This suggests that we might require that the set of conclusive reasons be closed under contraposition: whenever a proposition D is a conclusive reason for another proposition E, ~E is also a conclusive reason for ~D. If we impose this constraint on OSCAR then he will reason correctly in this case.

Humans handle this case differently. They use conditional reasoning to get the same result. If S is a conclusive reason for ~Q then by conditional reasoning they can obtain (S⊃~Q) even when S is not believed. Then once they reason (like OSCAR) from B to R to Q, they can go on to ~S, and then the rule (NEG-COL-DEFb) will lead to the retraction of Q, R, and ~S, and the addition of {<R,B>,<S,C>} to *rebut*. But OSCAR cannot reason this way until we give him conditional reasoning.

The main case in which OSCAR goes badly wrong is with collective undercutting defeat. He cannot handle this at all. Recall that a simple example of this occurs when Jones says "Smith is unreliable" and Smith says "Jones is unreliable." In such a case, we should not believe either. Suppose we take P to be a prima facie reason for Q, and Q to be a conclusive reason for ~(R→S), and in turn take R to be a prima facie reason for S and S to be a conclusive reason for ~(P→Q):

P- - ->Q-----> ~(R→S)
R- - ->S-----> ~(P→Q)

Then giving OSCAR the input {P,R} puts him into an infinite loop. I do not believe that this can be resolved without conditional reasoning. The difficulty can be seen by comparing collective undercutting defeat with collective rebutting defeat. The presence of collective rebutting defeat is signalled by the appearance of an explicit contradiction as the next step a reasoner would otherwise take. This hits the reasoner squarely in the face and cannot be ignored. But collective undercutting defeat can be much more subtle. We might have a long chain of reasoning like that in Figure 8, where all the steps are defeasible. This should be a case of collective undercutting defeat, but to discover such collective defeat the reasoner may have to trace his reasoning back arbitrarily far, in this case, all the way back to p_1 and q_1. We do not

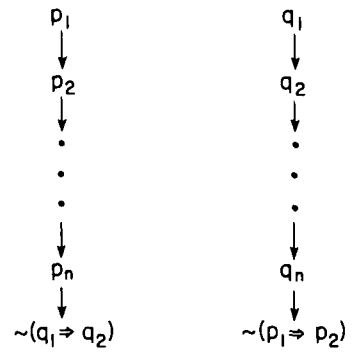

Figure 8.

want to require that OSCAR check for this every time he makes an inference, because the overhead would be immense and his reasoning would be slowed to a crawl.

I suggest that the way this is handled in human beings is that we do not usually worry about collective undercutting defeat, and we only take the inferences involved to be defeated if we happen upon it. We do not automatically check for it. Happening upon it consists of adopting certain conditionals. Specifically, we can use a rule something like the following:

> Where A is a prima facie reason for B and P is a prima facie reason for Q, if you come to believe B on A and Q on P, and also the conditionals $[B \rightarrow \sim(P \rightarrow Q)]$ and $[Q \rightarrow \sim(A \rightarrow B)]$, take $<Q,P>$ and $<B,A>$ to be subject to collective undercutting defeat.

However, such a rule can only be implemented in a system that includes conditional reasoning, so there is no way to build it into the present version of OSCAR.

I have not implemented the "self-defeating" constraint that I used to resolve the paradoxes of defeasible reasoning. Again, we do not want OSCAR to have to continually check whether his arguments are self-defeating, because that requires backtracking arbitrarily far. I suggest that this too is best handled in terms of conditional reasoning, but I will not go into it here.

OSCAR draws all possible conclusions from his input. We have, in effect, taken him to be interested in everything. This has the consequence that we must be careful what reason schemes we give him. For instance, if we allow him to use addition:

> p is a conclusive reason for '(p v q)'

then his reasoning can never stop. He will just go on inferring longer and longer disjunctions. Thus to test OSCAR I have had to be judicious in my choice of reason schemes. To rectify this we must incorporate interest-driven

reasoning. My next project is to design a system of interest-driven deductive reasoning, incorporating subsidiary arguments and conditionalization. Then I will integrate that into the present OSCAR to create an interest-driven system that does both defeasible and deductive reasoning in full generality.

REFERENCES

Chisholm, R.M. (1957). *Perceiving*. Ithaca, NY: Cornell University Press.

Chisholm, R.M. (1966). *Theory of knowledge*. Englewood Cliffs, NJ: Prentice-Hall.

Chisholm, R.M. (1977). *Theory of knowledge* (2nd ed.).Englewood Cliffs, NJ: Prentice-Hall.

Doyle, J. (1979). A truth maintenance system. *Artificial Intelligence, 12*, 231-272.

Harman, G. (1984). Positive versus negative undermining in belief revision. *Nous, 18*, 39-49.

Harman, G. (1986). *Change in view*. Cambridge, MA: MIT Press.

Kyburg, H., Jr. (1961). *Probability and the logic of rational belief*. Middletown, CT: Wesleyan University Press.

McCarthy, J. (1980). Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence, 13*, 27-39.

McDermott, D., & Doyle, J. (1980). Non-monotonic logic I. *Artificial Intelligence, 13*, 41-72.

Pollock, J.L. (1967). Criteria and our knowledge of the material world. *Philosophical Review, 76*, 28-62.

Pollock, J.L. (1970). The structure of epistemic justification. *American Philosophical Quarterly*. (Monograph series 4: 62-78).

Pollock, J.L. (1974). *Knowledge and justification*. Princeton: Princeton University Press.

Pollock, J.L. (1976). *Subjunctive reasoning*. Dordrecht: Reidel.

Pollock, J.L. (1983). Epistemology and probability. *Synthese, 55*, 231-252.

Pollock, J.L. (1984). A solution to the problem of induction. *Nous, 18*, 423-462.

Pollock, J.L. (1986). *Contemporary theories of knowledge*. Totowa, NJ: Rowman and Allanheld.

Pollock, J.L. (in press). How to build a person. *Philosophical Perspectives, 1*.

Pollock, J.L. (in preparation). *Nomic probability and the foundations of induction*.

Reiter, R. (1978). On reasoning by default. *Theoretical Issues in Natural Language Processing-2*, 210-218.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence, 13*, 81-132.