Chapter 3
The Structure of Defeasible Reasoning

1. Reasons and Defeaters

It was illustrated in chapter two that most rational thought involves reasoning that is *defeasible*, in the sense that the reasoning can lead not only to the adoption of new beliefs but also to the retraction of previously held beliefs. This chapter investigates the logical structure of defeasible epistemic reasoning. Defeasible practical reasoning will be investigated in chapter 7. The concern of this chapter is the semantics of defeasible reasoning (construed broadly, as in chapter one). Chapter 4 will turn to the procedural issue of how a rational agent can perform defeasible reasoning.

Reasoning proceeds by constructing arguments, where *reasons* provide the atomic links in arguments. *Conclusive reasons* are reasons that are not defeasible. Conclusive reasons logically entail their conclusions. Those that are not conclusive are *prima facie reasons*. Prima facie reasons create a presumption in favor of their conclusion, but it can be defeated. A reason will be encoded as an ordered pair $\langle \Gamma, p \rangle$, where Γ is the set of premises of the reason and p is the conclusion. Considerations that defeat prima facie reasons are *defeaters*. The simplest kind of defeater for a prima facie reason $\langle \Gamma, p \rangle$ is a reason for denying the conclusion. Let us define '$\neg$' as follows: if for some θ, $\varphi = \ulcorner \sim \theta \urcorner$, let $\neg\varphi = \theta$, and let $\neg\varphi = \ulcorner \sim \varphi \urcorner$ otherwise. Then we define:

> If $\langle \Gamma, p \rangle$ is a prima facie reason, $\langle \Lambda, q \rangle$ is a *rebutting defeater* for $\langle \Gamma, p \rangle$ iff $\langle \Lambda, q \rangle$ is a reason and $q = \ulcorner \neg p \urcorner$.

Prima facie reasons for which the only defeaters are rebutting defeaters would be analogous to normal defaults in default logic [Reiter 1980]. Experience in using prima facie reasons in epistemology indicates that there are no such prima facie reasons. Every prima facie reason has associated undercutting defeaters, and these are the most important kinds of defeaters for understanding any complicated reasoning. This is illustrated in chapter 2, and more fully in Pollock [1974, 1986, and 1990]. Undercutting defeaters attack a prima facie reason without attacking its conclusion. They accomplish this by attacking the connection between the premises and the conclusion. For instance, $\ulcorner x$ looks

red to me⌝ is a prima facie reason for an agent to believe ⌜x is red⌝. But if I know not only that x looks red but also that x is illuminated by red lights and red lights can make things look red when they are not, then it is unreasonable for me to infer that x is red. Consequently, ⌜x is illuminated by red lights and red lights can make things look red when they are not⌝ is a defeater, but it is not a reason for thinking that x is not red, so it is not a rebutting defeater. Instead, it attacks the connection between ⌜x looks red to me⌝ and ⌜x is red⌝, giving us a reason for doubting that x wouldn't look red unless it were red. ⌜P wouldn't be true unless Q were true⌝ is some kind of conditional, and I will symbolize it as ⌜P » Q⌝. If $\langle\Gamma,p\rangle$ is a prima facie reason, then where $\Pi\Gamma$ is the conjunction of the members of Γ, any reason for denying ⌜$\Pi\Gamma$ » p⌝ is a defeater. Thus I propose to characterize undercutting defeaters as follows:

> If $\langle\Gamma,p\rangle$ is a prima facie reason, $\langle\Lambda,q\rangle$ is an *undercutting defeater* for $\langle\Gamma,p\rangle$ iff $\langle\Lambda,q\rangle$ is a reason and $q =$ ⌜~($\Pi\Gamma$ » p)⌝.

It will be convenient to abbreviate ⌜~(P » Q)⌝ as ⌜(P ⊗ Q)⌝. I will henceforth represent undercutting defeaters as reasons for ⌜($\Pi\Gamma$ ⊗ p)⌝.

2. Arguments and Inference Graphs

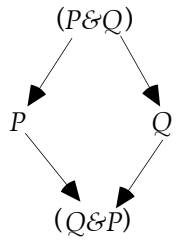
Reasoning starts with premises that are input to the reasoner. (In human beings, they are provided by perception.) The input premises comprise the set *input*. The reasoner then makes inferences (some conclusive, some defeasible) from those premises using reason schemas. *Arguments* are structures recording the reasoning that is performed. It was urged in chapter two that reasoning begins with perceptual states rather than beliefs about perceptual states. Accordingly, the arguments that summarize our reasoning must be viewed as structures consisting of mental states, some of which are perceptual states but most of which are beliefs. This is a somewhat unusual construal of arguments, which are usually taken to be structures of propositions rather than structures of mental states, but this construal is forced upon us by the facts about perception. It is occasionally awkward to view arguments in this way, however. When the constituents of arguments are belief states, we can just as well represent the beliefs in terms of the propositions believed, and doing so makes the arguments look more familiar.

The simplest arguments are *linear*. These can be viewed as finite sequences of mental states or propositions, each of which is either a member of *input* or inferable from previous members of the sequence in accordance with some reason schema. The order in the sequence

represents the order in which the inferences are made. Some aspects of that ordering may be inessential to the logical structure of the reasoning. Consider, for example, the following two arguments:

1. (*P&Q*)	1. (*P&Q*)
2. *P* from 1	2. *Q* from 1
3. *Q* from 1	3. *P* from 1
4. (*Q&P*) from 2 and 3.	4. (*Q&P*) from 2 and 3.

These represent different orders in which the inferences occur, but the dependency relations in the arguments are the same and could be represented more perspicuously as a graph:



Which of these structures we regard as the argument depends upon what we take the argument to be encoding. If we interpret the argument as encoding an actual sequence of inferences, then it will have the first (sequential) form. If instead we take it as encoding the dependency relations, it will have the second (graphical) form. Both structures are useful for different purposes, so I propose to refer to the sequential structures as *arguments* and the graphical structures as *inference graphs*. We can think of reasoning as a process that builds inference graphs, and the process of construction is recorded (or displayed) in arguments. For many purposes, arguments and inference graphs are interchangeable.

When a reasoner reasons, it is natural to regard it as producing a number of different arguments aimed at supporting different conclusions. However, we can combine all of the reasoning into a single inference graph that records the overall state of the reasoner's inferences, showing precisely what inferences have been made and how inferences are based upon one another. This comprehensive inference graph will provide the central data structure used in evaluating a reasoner's beliefs. Accordingly, we can think of the function of reasoning to be that of building the inference graph. I will return to this point later.

Linear reasoning is a particularly simple form of reasoning in which each conclusion drawn is either given as a member of *input* or inferred

from previous conclusions. Not all reasoning is linear. To see this, note that linear reasoning can only lead to conclusions that depend upon the members of *input*, but actual reasoning can lead to a priori conclusions like $(p \vee \sim p)$ or $((p \ \& \ q) \supset q)$ that do not depend upon anything. What makes this possible is *suppositional reasoning*. In suppositional reasoning we "suppose" something that we have not inferred from *input*, draw conclusions from the supposition, and then "discharge" the supposition to obtain a related conclusion that no longer depends upon the supposition. The simplest example of such suppositional reasoning is *conditionalization*. When using conditionalization to obtain a conditional $(p \supset q)$, we suppose the antecedent p, somehow infer the consequent q from it, and then discharge the supposition to infer $(p \supset q)$ independently of the supposition. Similarly, in *reductio ad absurdum* reasoning, to obtain $\neg p$ we may suppose p, somehow infer $\neg p$ on the basis of the supposition, and then discharge the supposition and conclude $\neg p$ independently of the supposition. Another variety of suppositional reasoning is *dilemma* (reasoning by cases).

If we are to encode suppositional reasoning in the inference graph, then the nodes of the inference graph must correspond to conclusions drawn *relative to particular suppositions*. To accomplish this, I will take the nodes of the inference graph to encode inferences to *sequents*, where a sequent is an ordered pair $\langle X,p\rangle$ consisting of a supposition (a set of propositions X) and a conclusion (a single proposition) p. If a node records the inference of a conclusion p relative to a supposition X, I will say that the node *supports p relative to the supposition X* or, alternatively, *supports the sequent* $\langle X,p\rangle$. If $X = \emptyset$, I will say simply that the node *supports p*. The inference relations between nodes are recorded in *inference links*. Where ν and η are nodes, $\langle \nu,\eta\rangle$ is an inference link iff ν was inferred from a set of nodes one of which was η. The *immediate inference ancestors* of a node are the nodes to which it is connected (from which it was inferred) by inference links. One complication calls for explicit mention. A reasoner might construct more than one argument supporting the same sequent. The nodes of the inference graph encode *inferences*, so there must be a separate node, each with its own inference links, for each argument supporting the sequent. This will allow us to associate unique strengths with nodes, regard different nodes as defeated by different defeaters, and so on. An *inference branch* is a finite sequence of nodes each of which is an immediate inference ancestor for the next. Let us say that η is an *inference ancestor* of ν iff there is an inference branch connecting ν to η. A node is a *pf-node* iff it represents a defeasible inference, in accordance with some prima facie reason. Let us say that μ is a *deductive ancestor* of ν iff μ is an inference ancestor of ν and the branch connecting them contains no pf-nodes. μ is a *nearest defeasible ancestor* of ν iff either (1) ν is a

pf-node and $\mu = \nu$ or (2) μ is a deductive ancestor of ν and μ is a pf-node.

We normally talk about propositions being believed or disbelieved, but we need some similar terminology for talking about sequents in general, so I will apply the term "belief" to sequents. Belief in propositions corresponds to belief in sequents having empty suppositions. Belief in a sequent with a nonempty supposition might be called "conditional belief". An agent's beliefs constitute only a subset of the conclusions represented in the agent's inference graph, because when a conclusion is defeated by another conclusion, it is still represented in the inference graph, but it may not be believed. A rational agent's beliefs are its undefeated conclusions.

Rules of inference can be viewed as rules for adding nodes to the agent's inference graph. One way of conceptualizing this is to think of rules of inference as clauses in the recursive definition of "inference graph". On this conception, an inference graph is any set of inference nodes that can be constructed by starting from the set *input* and accumulating nodes in accordance with the rules of inference. Viewing rules of inference in this way, the following are some obvious inference rules:

Input

If $p \in$ *input* and G is an inference graph, then for any supposition X, a new inference graph can be constructed by adding to G a node supporting $\langle X,p\rangle$ and letting the set of immediate inference ancestors of the new node be empty.

Supposition

If G is an inference graph and X is any finite set of propositions, then if $p \in X$, a new inference graph can be constructed by adding to G a node supporting $\langle X,p\rangle$, and letting the set of immediate inference ancestors of the new node be empty.

Reason

If G is an inference graph containing nodes $\alpha_1,...,\alpha_n$ supporting each of $\langle X,p_1\rangle,...,\langle X,p_n\rangle$, and $\langle\{p_1,...,p_n\},q\rangle$ is a reason (either conclusive or prima facie), then a new inference graph can be constructed by adding to G a node supporting $\langle X,q\rangle$, and letting the set of immediate inference ancestors of the new node be $\{\alpha_1,...,\alpha_n\}$.

Conditionalization

If G is an inference graph containing a node α supporting $\langle X \cup \{p\}, q\rangle$ then a new inference graph can be constructed by adding to G a node supporting $\langle X, (p \supset q)\rangle$, and letting the set of immediate inference ancestors of the new node be {α}.

Dilemma

If G is an inference graph containing a node α supporting $\langle X, (p \vee q)\rangle$, a node β supporting $\langle X \cup \{p\}, r\rangle$, and a node γ supporting $\langle X \cup \{q\}, r\rangle$, then a new inference graph can be constructed by adding to G a node supporting $\langle X, r\rangle$, and letting the set of immediate inference ancestors of the new node be {α,β,γ}.

Other rules of inference graph formation could be included as well, but these will suffice for illustrative purposes.

A distinction must be made between two different kinds of suppositions and suppositional reasoning. In *factual* suppositional reasoning, we suppose that something *is* the case and then reason about what else is the case. In *counterfactual* suppositional reasoning, we make a supposition of the form, "Suppose it *were* true that *P*", and then reason about what *would* be the case. Both kinds of reasoning appear to support forms of conditionalization, but the inferred conditionals are different. Conditionalization from counterfactual suppositions yields counterfactual conditionals, whereas conditionalization from factual suppositions yields only material conditionals. These two kinds of suppositional reasoning work in importantly different ways. In factual suppositional reasoning, because we are supposing that something *is* the case, we should be able to combine the supposition with anything we have already concluded to be the case. Counterfactual suppositions, on the other hand, override earlier conclusions and may require their retraction within the supposition. Counterfactual suppositional reasoning is extremely interesting, but throughout this book, I will appeal only to factual suppositional reasoning. For such reasoning, the following inference rule is reasonable:

Foreign Adoptions

If G is an inference graph containing a node α supporting $\langle X, p\rangle$, and $X \subseteq Y$, then a new inference graph can be constructed by adding to G a node supporting $\langle Y, p\rangle$ and letting the immediate inference ancestors of the new node be the same as the immediate inference ancestors of α.

This rule is reasonable because in supposing that something is true, we should be able to combine it with anything else we are already

justified in believing to be true. This rule would not be reasonable for counterfactual suppositional reasoning.

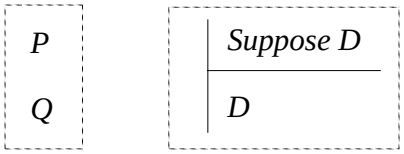
3. Defeat among Inferences—Uniform Reasons

The concept of epistemic justification that is of relevance to this investigation concerns belief updating. Justified beliefs are those mandated by the rules for belief updating. What an agent is justified in believing is a function of both what *input* premises have been supplied by the perceptual systems and how far the agent has gotten in its reasoning. A necessary condition for a belief to be jusitified is that the agent has engaged in reasoning that produced an argument supporting the belief, but that is not a sufficient condition because the agent may also have produced an argument that defeats the first argument.

The agent's reasoning is encoded in the inference graph, the nodes of which correspond to inferences. Defeat is both a relation between nodes of the inference graph and a monadic property of nodes. A node is defeated (has the monadic property) just in case it stands in appropriate defeat relations to other nodes. A justified belief is one supported by an undefeated node of the inference graph.

To complete this characterization of epistemic justification, we need an account of the defeat relations and a characterization of when those defeat relations render a node defeated. A general treatment of defeat relations requires us to take account of the fact that reasons differ in strength; some are better than others. If we have a reason for p and a reason for $\neg p$, but the latter is stronger than the former, then it wins the competition and we should believe $\neg p$. A general theory of reasoning requires us to talk about the strengths of reasons and how those strengths affect interactions between reasons. However, before addressing this complicated issue, consider how defeat between nodes could be analyzed if all reasons were of the same strength. This is *The Assumption of Uniform Reasons*.

One node of the inference graph defeats another by supporting a defeater for it. Recall, however, that in suppositional reasoning, different steps of reasoning may depend upon different suppositions. If a node η of the inference graph results from applying a prima facie reason $\langle \Gamma, p \rangle$, and a node σ supports a defeater for this prima facie reason, this does not automatically guarantee that σ defeats η because they may support their conclusions relative to different suppositions. Consider, for instance, the following pair of arguments, where P is a prima facie reason for Q, and D is a defeater for this prima facie reason:



These could be represented by the inference graph diagrammed in figure 1. The defeater D is introduced as a mere supposition, and that supposition is not included in the suppositions made in η. Clearly we should not be able to defeat an inference just by *supposing* a defeater that has no independent justification. It seems that the support of a defeater by a node whose supposition is X should defeat only a defeasible

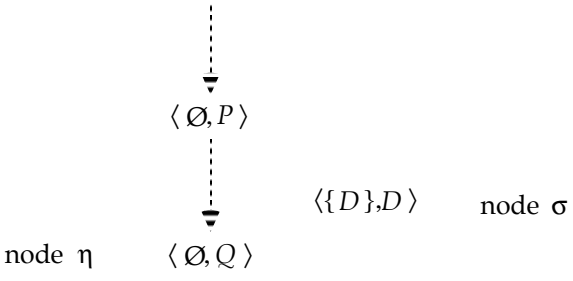


Figure 1. Supposing a defeater

inference that is made relative to a supposition that includes X. Accordingly, on the assumption of uniform reasons we can define:

A node σ *rebuts* a node η iff:
(1) η is a pf-node supporting some proposition q relative to a supposition Y; and
(2) σ supports $\neg q$ relative to a supposition X, where $X \subseteq Y$.

A node σ *undercuts* a node η iff:
(1) η is a pf-node supporting some proposition q relative to a

supposition Y; where $p_1,...,p_k$ are the propositions supported by its immediate ancestors; and

(2) σ supports $((p_1$ & ... & $p_k) \otimes q)$ relative to a supposition X where $X \subseteq Y$.

A node σ *defeats* a node η iff σ either rebuts or undercuts η.

Now let us consider how this account of defeat relations must be modified to accommodate reasons of varying strength.

4. Taking Strength Seriously

At this point, I will relax the simplifying assumption that all reasons are of the same strength.

Measuring Strength

If we are to take strength seriously, we must have some way of measuring it. One way is to compare reasons with a set of standard equally good reasons that have numerical values associated with them in some determinant way. I propose to do that by taking the set of standard reasons to consist of instances of the statistical syllogism. Recall from chapter two that this principle is formulated as follows:

The Statistical Syllogism:

If $r > 0.5$ then $\ulcorner \text{prob}(F/G) \geq r$ & $Gc \urcorner$ is a prima facie reason for $\ulcorner Fc \urcorner$, the strength of the reason being a monotonic increasing function of r.

Consequently, for any proposition p, we can construct a standardized argument for $\neg p$ on the basis of the pair of suppositions $\ulcorner \text{prob}(F/G) \geq r$ & $Gc \urcorner$ and $\ulcorner (p \equiv \sim Fc) \urcorner$:

1. Suppose $\text{prob}(F/G) \geq r$ & Gc.
2. Suppose $(p \equiv \sim Fc)$.
3. Fc from 1.
4. $\neg p$ from 2,3.

where the strength of the argument is a function of r. If X is a prima facie reason for p, we can measure the strength of this prima facie reason in terms of that value of r such that the conflicting argument from the suppositions $\ulcorner \text{prob}(F/G) \geq r$ & $Gc \urcorner$ and $\ulcorner (p \equiv \sim Fc) \urcorner$ exactly counteracts it. We could take r itself to be the measure of the strength

of the reason, but a somewhat more convenient measure is $2\cdot(r - 0.5)$. This has the convenient consequence that the strength of an instance of statistical syllogism in which $r = 0.5$ is 0, and strengths are normalized to 1.0. So my proposal is:

> If X is a prima facie reason for p, the strength of this reason is $2\cdot(r - 0.5)$ where r is that real number such that an argument for $\neg p$ based upon the suppositions $\ulcorner \text{prob}(F/G) \geq r \ \&\ Gc \urcorner$ and $\ulcorner (p \equiv \sim Fc) \urcorner$ and employing the statistical syllogism exactly counteracts the argument for p based upon the supposition X.

For instance, in a context in which a prima facie reason of minimal acceptable strength corresponds to an instance of the statistical syllogism in which $r = 0.95$, it follows that prima facie reasons must have a strength ≥ 0.9.

Conclusive reasons logically guarantee the truth of their conclusions given the truth of the premises, so there can be no accompanying attenuation in strength of justification. We can capture this by taking them to have strength 1.0.

I take it that this way of measuring the strengths of reasons is very natural. However, it has an important consequence that deserves emphasis. Instances of the statistical syllogism can be linearly ordered by strength, because that is a function of the real number r. Other reasons can be compared to instances of the statistical syllogism having the same strength, so it follows that all reasons can be linearly ordered by strength. Although this consequence is unsurprising, it does conflict with some recent proposals regarding nonmonotonic reasoning that assume the ordering of reasons by strength to be only a partial ordering.

Generic Bayesianism

Given a measure of the strengths of reasons, what are we to do with it? Strengths are important in deciding whether a reason is strong enough to justify a belief. It was indicated in chapter 1 that the strength of reason required to justify a conclusion is a function of our degree of interest in the conclusion. (This will be addressed further in chapter four.) However, there is a residual problem. Although individual reasons may be sufficiently strong to justify their conclusions in a one-step argument, how do we determine the degree of justification of the conclusion of a complex argument that combines inferences using a number of such reasons?

It is useful to distinguish between the degree of support of a conclusion and its degree of justification. Whether a conclusion is justified depends not just on how strong the arguments are that support it but also on whether those arguments are defeated. My present concern is with the strengths of arguments that combine inferences employing a number of reasons that are of less than unit strength. It is often supposed that, in such an argument, each inference attenuates the strength of the conclusion, and so, although each reason by itself is sufficiently strong, the degree of support of the ultimate conclusion may be too weak to justify believing it.

This supposition is usually coupled with a probabilistic model of reasoning according to which reasons make their conclusions probable to varying degrees, and the ultimate conclusion is justified only if it is made sufficiently probable by the cumulative reasoning. I will refer to this theory as *generic Bayesianism*.

According to generic Bayesianism, our epistemic attitude towards a proposition should be determined by its probability. It will generally be necessary to compute such probabilities in order to determine the degree of justification of a belief. Some kinds of deductive inference can be applied "blindly" without going through such calculations, but only when the inferences are guaranteed to preserve probability. Let us say that an inference rule

$$\frac{P_1,\ldots,P_n}{Q}$$

is *probabilistically valid* just in case it follows from the probability calculus that prob(Q) ≥ the minimum of the prob(P_i)'s. For the generic Bayesian, inference rules can be applied blindly, obviating the need for probability calculations, only if they are probabilistically valid.

If P logically entails Q, then it follows from the probability calculus that prob(Q) ≥ prob(P), and hence the generic Bayesian is able to conclude that the degree of justification for Q is as great as that for P. Thus deductive inferences from single premises can proceed blindly. However, this is not equally true for entailments requiring multiple premises. Specifically, it is not true in general that if {P,Q} entails R, then prob(R) ≥ the minimum of prob(P) and prob(Q). For instance, {P,Q} entails (P&Q), but prob(P&Q) may be less than either prob(P) or prob(Q).

Many of our most cherished inference rules, including *modus ponens*, *modus tollens*, and adjunction, turn out to be probabilistically invalid. No inference rule that proceeds from mutliple premises and uses them all essentially can be probabilistically valid. This is extremely counter-intuitive. It means that a reasoner engaging in Bayesian updating is precluded from drawing deductive conclusions from its reasonably held beliefs. For instance, consider an engineer who is designing a bridge. She will combine a vast amount of information about material strength, weather conditions, maximum load, costs of various construction techniques, and so forth, to compute the size a particular girder must be. These various bits of information are, presumably, independent of one another, so if the engineer combines 100 pieces of information, each with a probability of 0.99, the conjunction of that information has a probability of only $.99^{100}$, which is approximately 0.366. According to generic Bayesianism, she would be precluded from using all of this information simultaneously in an inference—but then it would be impossible to build bridges.

As a description of human reasoning, this seems clearly wrong. Once one has arrived at a set of conclusions, one does not hesitate to make further deductive inferences from them. But an even more serious difficulty for generic Bayesianism is that the theory turns out to be self-defeating; if the theory were correct, it would be impossible to perform the very calculations required by the theory for determining whether a belief ought to be held. This arises from the fact that Bayesian updating requires a reasoner to decide what to believe by computing probabilities. The difficulty is that the probability calculations themselves cannot be performed by a Bayesian reasoner. To illustrate the difficulty, suppose the reasoner has the following beliefs:

(4.1) $\text{prob}(P \vee Q) = \text{prob}(P) + \text{prob}(Q) - \text{prob}(P \& Q)$
$\text{prob}(P) = 0.5$
$\text{prob}(Q) = 0.49$
$\text{prob}(P \& Q) = 0.$

From this we would like the reasoner to compute that $\text{prob}(P \vee Q) = 0.99$, and perhaps go on to adopt $(P \vee Q)$ as one of its beliefs. However, generic Bayesianism cannot accommodate this. The difficulty is that the use of (4.1) in a computation is an example of a "blind use" of a deductive inference, and as such it is legitimate only if the inference is probabilistically valid. To determine the probabilistic validity of this inference, it must be treated on a par with all the other inferences performed by the reasoner. Although the premises are about probabilities, they must also be assigned probabilities ("higher-order probabil-

ities") to be used in their manipulation. Viewing (4.1) in this way, we find that although the four premises do logically entail the conclusion, the inference is not probabilistically valid for the same reason that *modus ponens*, *modus tollens*, and adjunction fail to be probabilistically valid. It is an inference from a mutliple premise set, and despite the entailment, the conclusion can be less probable than any of the premises.

How serious this difficulty is depends upon the probabilities of the four probabilistic premises. The first premise is a necessary truth, so it must have probability 1. If the other premises also have probability 1, then it follows that the conclusion has probability 1 and so the inference is probabilistically valid after all. (Similarly, *modus ponens* is probabilistically valid for the special case in which the premises have probability 1.) Can we assume, however, that a belief like ⌜prob(P) = 0.5⌝ automatically has probability either 1 or 0? Unless we are talking about Carnapian logical probability [Carnap 1950, 1952], that would seem totally unreasonable, and as such logical probabilities are insensitive to empirical facts, they are not appropriate for use in belief updating. The inescapable conclusion is that a Bayesian reasoner cannot perform the very calculations that are required for Bayesian reasoning.

It seems initially that there remains one possible avenue of escape for the generic Bayesian. If the set $\{P_1,...,P_n\}$ entails Q, the inference from $P_1,...,P_n$ to Q will not usually be probabilistically valid, but the inference from the *conjunction* $(P_1\&...\&P_n)$ to Q will be probabilistically valid. If in addition to believing each of the premises in (4.1), the reasoner also believes their conjunction, then it can validly infer that prob($P\vee Q$) = 0.99. The reasoner cannot validly infer the conjunction of the premises from the individual premises, because adjunction is probabilistically invalid, but if in addition to knowing that the premises in (4.1) are highly probable, the reasoner also knows the conditional probabilities of each on conjunctions of the others, it can use those conditional probabilities to compute the probability of the conjunction of the premises by using the following principle:

(4.2) $\quad \text{prob}(P\&Q) = \text{prob}(P/Q)\cdot\text{prob}(Q).$

This is "the conditional probability strategy". Unfortunately, this strategy is subject to an overwhelming difficulty: even if the reasoner had the requisite conditional probabilities, it would be unable to use them to perform the computation in (4.2). That computation would involve reasoning as follows:

$\text{prob}(P\&Q) = \text{prob}(P/Q)\cdot\text{prob}(Q).$
$\text{prob}(P/Q) = \alpha$
$\text{prob}(Q) = \beta$

$$\text{prob}(P\&Q) = \alpha\cdot\beta$$

The difficulty we encounter is precisely the same as the one we set out to solve in the first place: this is an inference from multiple premises, and as such is probabilistically invalid. To turn it into a valid inference, the reasoner would have to know the probability of the conjunction of the premises and make the inference from that conjunction instead. It was for this purpose that the conditional probability strategy was proposed initially. Thus rather than solving the problem, the strategy leads to an infinite regress.

The only conclusion that can be drawn from all of this is that generic Bayesianism is incoherent as a theory of belief updating. The probability calculations required of the reasoner proceed via deductive inferences that are not probabilistically valid, and hence the Bayesian reasoner is precluded from making the very calculations it needs to determine degrees of justification.

Generic Bayesianism is based upon the intuition that a proposition should be believed only if it is highly probable. If generic Bayesianism is logically incoherent, why is this intuition so compelling? The answer seems to be that the English word "probable" is ambiguous. We must distinguish between *statistical probability* and *epistemic probability*. Statistical probability is concerned with *chance*. We are using statistical probability when we talk about how likely it is to rain tomorrow, or about the probability of being dealt a royal flush in poker. Epistemic probability is concerned with the degree of justification of a belief. We are referring to epistemic probability when we conclude that the butler probably did it. All that means is that there is good reason to think the butler did it. It is certainly true that a belief is justified iff its epistemic probability is high; that is just what we mean by high epistemic probability. The lesson to be learned from the previous discussion is that rules like *modus ponens* and adjunction preserve high epistemic probability, and hence epistemic probability cannot be quantified in a way that conforms to the probability calculus. This should not be particularly surprising. There was never really any reason to expect epistemic probability to conform to the probability calculus. That is a calculus of statistical probabilities, and the only apparent connection between statistical and epistemic probability is that they share the same ambiguous name. It should have been obvious from the start that high epistemic probability is preserved by ordinary inference rules, and hence epistemic probability does not conform to the probability calculus.

The Weakest Link Principle

I have argued that generic Bayesianism must be rejected—belief updat-

ing cannot be performed by exclusively probabilistic methods. My proposal is that the degree of support for a conclusion should instead be computed in terms of the *Weakest Link Principle*, according to which a deductive argument is as good as its weakest link. More precisely:

> The degree of support of the conclusion of a deductive argument is the minimum of the degrees of support of its premises.

The simplest reason for favoring this principle is that it seems to be the only alternative to generic Bayesianism that is not completely ad hoc. Any other theory owes us an account of how the strength of an argument decreases as we add inferences from new premises. The only obvious account of that is the Bayesian account, but we have seen that it must be rejected.

There is also a strong argument in favor of the weakest link principle. This turns upon the fact that the objections to the Bayesian account can be applied more generally to any account that allows the strength of an argument to be less than its weakest link. On any such account, multi-premise inference rules like *modus ponens* and adjunction will turn out to be invalid, but then it seems unavoidable that the theory will be self-defeating in the same way as the Bayesian theory—by making it impossible for the reasoner to compute the degrees of support of its conclusions.

My defense of the weakest link principle has consisted of an attack on the alternatives, but there is also a very natural objection to the weakest link principle itself: faced with a long argument proceeding from multiple premises, we likely apt to view it with more suspicion than a simpler argument on the grounds that there is more that could go wrong with it. This sounds like a very Bayesian observation. However, its significance becomes less clear when we reflect that we are also inclined to regard complex, purely deductive arguments (proceeding from a priori premises) with suspicion, also on the grounds that there is more that could go wrong with them than with simple arguments. This certainly cannot be taken to show that the degree of support attaching to the conclusion of a deductive argument diminishes with the complexity of the argument (and correlatively that rules like *modus ponens* are invalid in purely deductive arguments), so it is doubtful that the analogous observation should have that consequence for arguments with contingent premises. I am inclined to think that what these observations really illustrate is the supervision of the reflexive reasoner rather than the built-in structure of the planar reasoner.

The above formulation of the weakest link principle applies only to deductive arguments, but we can use it to obtain an analogous principle for defeasible arguments. If P is a prima facie reason for Q,

then we can use conditionalization to construct a simple defeasible argument for the conclusion $(P \supset Q)$, and this argument turns upon no premises:

| Suppose P
|---
| Then (defeasibly) Q.

Therefore, $(P \supset Q)$.

Because this argument has no premises, the degree of support of its conclusion should be a function of nothing but the strength of the prima facie reason. Next, notice that any defeasible argument can be reformulated so that prima facie reasons are used only in subarguments of this form, and then all subsequent steps of reasoning are deductive. The conclusion of the defeasible argument is thus a deductive consequence of members of *input* together with a number of conditionals justified in this way. By the weakest link principle for deductive arguments, the degree of support of the conclusion should then be the minimum of (1) the degrees of justification of the members of *input* used in the argument and (2) the strengths of the prima facie reasons.

Input consists of states like "That looks red to me", from which one can infer defeasibly, "That is red". Because something can look more or less clearly red, and that can affect the justification of the conclusion, we must assign differing "strengths" to the *input* states, depending upon how clearly the object looks red. These strengths must be factored into the computation of the degree of support for the conclusion of a defeasible argument. This yields *The Weakest Link Principle for Defeasible Arguments*:

> The degree of support of the conclusion of a defeasible argument is the minimum of the strengths of the prima facie reasons employed in it and the strengths of the *input* states to which it appeals.

The problem of computing degrees of support is thus computationally simple. Sometimes it will be convenient to talk about the *strength of the argument* as being the degree of support of its conclusion. Each node of the inference graph will be assigned a unique strength—the strength of the argument supporting it, computed in accordance with the weakest link principle.

The Accrual of Reasons

If we have two independent reasons for a conclusion, does that make

the conclusion more justified than if we had just one? It is natural to suppose that it does, but upon closer inspection that becomes unclear. Cases that seem initially to illustrate such accrual of justification appear upon reflection to be better construed as cases of having a single reason that subsumes the two separate reasons. For instance, if Jones tells me that the president of Slobovia has been assassinated, that gives me a reason for believing it; and if Smith tells me that the president of Slobovia has been assassinated, that also gives me a reason for believing it. Surely, if they both tell me the same thing, that gives me a better reason for believing it. However, there are considerations indicating that my reason in the latter case is not simply the conjunction of the two reasons I have in the former cases. Reasoning based upon testimony is a straightforward instance of the statistical syllogism. We know that people tend to tell the truth, and so when someone tells us something, that gives us a prima facie reason for believing it. This turns upon the following probability being reasonably high:

(1) prob(p is true / S asserts p).

When we have the concurring testimony of two people, our degree of justification is not somehow computed by applying a predetermined function to the latter probability. Instead, it is based upon the quite distinct probability

(2) prob(p is true / S_1 asserts p and S_2 asserts p and $S_1 \neq S_2$).

The relationship between (1) and (2) depends upon contingent facts about the linguistic community. We might have one community in which speakers tend to make assertions completely independently of one another, in which case (2) > (1); and we might have another community in which speakers tend to confirm each other's statements only when they are fabrications, in which case (2) < (1). Clearly our degree of justification for believing p will be different in the two linguistic communities. It will depend upon the value of (2), rather than being some function of (1).

All examples I have considered that seem initially to illustrate the accrual of reasons turn out in the end to have this same form. They are all cases in which we can estimate probabilities analogous to (2) and make our inferences on the basis of the statistical syllogism rather than on the basis of the original reasons. Accordingly, I doubt that reasons do accrue. If we have two separate undefeated arguments for a conclusion, the degree of justification for the conclusion is simply

the maximum of the strengths of the two arguments. This will be my assumption.

Defeat among Inferences

One of the most important roles of the strengths of reasons lies in deciding what to believe when one has conflicting arguments for q and $\neg q$. It is clear that if the argument for q is *much* stronger than the argument for $\neg q$, then p should be believed, but what if the argument for q is just slightly stronger than the argument for $\neg q$? It is tempting to suppose that the argument for $\neg q$ should at least attenuate our degree of confidence in q, in effect lowering its degree of justification. But upon further reflection, I am inclined to think that this is false. Otherwise, if we acquired a second argument for $\neg q$, it would face off against a weaker argument for q and so be better able to defeat it. But that is tantamount to taking the two arguments for $\neg q$ to result in greater justification for that conclusion, and that is just the principle of accrual. So it seems that if we are to reject the latter principle, then we should also conclude that arguments that face weaker conflicting arguments are not thereby diminished in strength. There are cases that initially appear to be counterexamples to this. For instance, if Jones, whom I regard as highly reliable, tells me that the president of Slobovia has been assassinated, this may justify me in believing that the president of Slobovia has been assassinated. If Smith, whom I regard as signficantly less reliable, denies this, his denial may be insufficient to defeat my belief in the assassination, but surely Smith's denial should render me less confident of my belief. However, this example is analogous to those that appeared at first to illustrate the accrual of justification. What is happening is that contingent information is leading me to believe that the probability of an assertion's being true is lowered by its being denied by another at least somewhat reliable source, and this constitutes a subproperty defeater for the application of the statistical syllogism that is involved in my reasoning in this case. Accordingly, the original argument (based simply on Jones' assertion) is defeated outright, and I must instead base my belief in the assassination on the lower probability of its having happened given that Jones reported it and Smith denied it.

In the light of the preceding considerations, I think it should be concluded that an argument for $\neg q$ defeats an argument for q only if it is at least as strong as the argument for q. Accordingly, we can charac-

terize rebutting defeat as follows:

A node α *rebuts* a node β iff:
(1) β is a pf-node of some strength ξ supporting some proposition q relative to a supposition Y;
(2) α is a node of strength η and supports $\neg q$ relative to a supposition X, where $X \subseteq Y$; and
(3) $\eta \geq \xi$.

How does strength affect undercutting defeat? In previous publications, I have assumed that it does not—that all that is required for undercutting defeat is that the undercutting defeater be justified. But this view becomes incoherent when we take account of the fact that justification must always be relativized to a degree of justification.[1] It seems apparent that any adequate account of justification must have the consequence that if a belief is unjustified relative to a particular degree of justification, then it is unjustified relative to any higher degree of justification. However, this obvious constraint is violated by the principle that all that is required for undercutting defeat is that the undercutting defeater be justified. Suppose we have a strong reason P for some conclusion Q and a weaker reason for $(P \otimes Q)$. Relative to a low degree of justification, the undercutting defeater $(P \otimes Q)$ would be justified, and so Q would be unjustified. But relative to a higher degree of justification, $(P \otimes Q)$ would be unjustified, and so Q would turn out to be justified. This is perverse. It seems undeniable that degrees of justification must enter into undercutting defeat in the same way as for rebutting defeat:

A node α *undercuts* a node β iff:
(1) β is a pf-node of some strength ξ supporting some proposition q relative to a supposition Y; where $p_1,...,p_k$ are the propositions supported by its immediate ancestors; and
(2) α is a node of strength η and supports $((p_1 \,\&\, ... \,\&\, p_k) \otimes q)$ relative to a supposition X where $X \subseteq Y$; and
(3) $\eta \geq \xi$.

It will be convenient to encode defeat relations between nodes of

[1] I was led to see this as a result of discussions with Stewart Cohen.

the inference graph in a new set of links, called *defeat links*. Where μ and ν are nodes of the inference graph, ⟨μ,ν⟩ is a defeat link iff μ is a pf-node and ν defeats it relative to some degree of justification.

5. Other Nonmonotonic Formalisms

Although in general outline the theory of defeasible reasoning presented predates the most familiar theories of nonmonotonic reasoning in AI, it will seem unfamiliar to many researchers in AI because of the past isolation of philosophy and AI from one other. Accordingly, it will be useful to compare this theory to those more familiar in AI. The comparison will be scattered through the next several sections. Because the AI theories will be unfamiliar to most philosophers, I will give a brief account of each.

Argument-Based Approaches

The theory of defeasible reasoning adumbrated in this book is an "argument-based" theory, in the sense that it characterizes defeasible consequence in terms of the interactions between the inference steps of all possible arguments that can be constructed from the given set *input* using a fixed set of prima facie reasons and defeaters. Other argument-based approaches to defeasibility can be found in the work of Loui [1987], Simari and Loui [1992], and Lin and Shoham [1990]. The work of Horty, Thomason, and Touretzky [1990] and Horty and Thomason [1990] can also be viewed as an argument-based theory of defeasible reasoning, where the arguments have a rather restricted form. However, these theories are all based upon rather simple conceptions of argument, confining their attention to linear arguments. None of the AI theories of nonmonotonic reasoning appears to be sensitive to the importance of suppositional reasoning, but suppositional reasoning is essential in any reasoner that is capable of performing deductive and defeasible reasoning simultaneously. In addition, the systems of Horty, Thomason, and Touretzky [1990], Horty and Thomason [1990], and Simari and Loui [1992] accommodate only rebutting defeaters.

Much of the work that lies within the argument-based approach adds a form of "specificity defeat" that is not part of the theory adumbrated in this book. The details of specificity defeat have been worked out differently by different authors, but the basic idea is that when two arguments have conflicting conclusions, if one of the arguments is based upon a "more specific" set of premises, it should take precedence

and the other argument should be regarded as defeated. Specificity defeat originated with Poole [1985, 1988], and represents a generalization of the subproperty defeaters that play so prominent a role in connection with the statistical syllogism. My own view is that the correctness of specificity defeat turns heavily upon the context in which it is being used. Horty, Thomason, and Touretzky [1990] apply it to defeasible inheritance hierarchies in which all nodes are logically simple. Such hierarchies can be viewed as networks of concepts related in the manner required for the application of the statistical syllogism, and the conclusions of such defeasible inheritance proceed in accordance with the statistical syllogism. So construed, specificity defeat amounts to subproperty defeat, and if it is formulated properly I have no objections to it. But in Horty and Thomason [1990] the defeasible inheritance hierarchies are extended so that the nodes can represent Boolean combinations of concepts (concepts constructed out of simple concepts using negation, conjunction, and disjunction). Here, there is a problem. As we saw in chapter 2, the principle of subproperty defeat must incorporate a projectibility constraint, and that has the effect of ruling out many disjunctions and negations. The result is that specificity defeat cannot be applied in general within such logically complex defeasible inheritance hierarchies. This difficulty has been completely overlooked in the literature.

Loui [1987] and Simari and Loui [1992] treat specificity defeat more generally, applying it to all defeasible reasoning, and not just defeasible reasoning in accordance with the statistical syllogism. This strikes me as entirely wrong. First, in cases in which the reasoning can be regarded as proceeding in accordance with the statistical syllogism, they will encounter the same projectibility problems as Horty and Thomason. Second, it was observed in chapter 2 that much defeasible reasoning cannot be regarded as proceeding in accordance with the statistical syllogism. For such reasoning, there is no reason to think that specificity defeat is a correct principle of reasoning. The only intuitive examples that have ever been given in support of it concern the statistical syllogism. The literature contains absolutely no examples illustrating its applicability elsewhere, and a conscientious effort on my part to construct such examples has ended in complete failure. It is for this reason that I reject specificity defeat as a general principle of defeasible reasoning.

Default Logic

In spirit, my own theory of defeasible reasoning seems close to Reiter's default logic [1980], with prima facie reasons and defeaters corresponding to Reiter's defaults. Defaults can be regarded as defeasible inference rules. Reiter writes defaults in the form $\ulcorner P{:}Q_1,\ldots,Q_n/R\urcorner$, the

interpretation being that P provides the premise for a default inference to the conclusion R, and any of $\sim Q_1,...,\sim Q_n$ will defeat the inference. Given a set D of defaults and a set W of premises, Reiter defines an *extension* of $\langle D,W\rangle$ to be a set E such that (1) $W \subseteq E$; (2) E is closed under deductive consequence; (3) for each default $\ulcorner P:Q_1,...,Q_n/R\urcorner$ in D, if $P \in E$ but none of the $\sim Q_i$ is in E, then R is in E; and (4) no proper subset E^* of E satisfies (1), (2), and the condition (3*) that for each default $\ulcorner P:Q_1,...,Q_n/R\urcorner$ in D, if $P \in E^*$ but none of the $\sim Q_i$ is in E, then R is in E^*. The extensions of pairs of defaults and premises are supposed to represent the sets of rational belief sets one could have if one were given those defaults and premises as one's starting point. This is rationality in the sense of "rationality in the limit", that is, what one could reasonably believe if one had performed all relevant reasoning.

Although default logic is, in some ways, very similar to the theory of defeasible reasoning, there are also profound differences between the two theories. First, prima facie reasons are supposed to be logical relationships between concepts. It is a necessary feature of the concept *red* that something's looking red to me gives me a prima facie reason for thinking it is red. (To suppose we have to discover such connections inductively leads to an infinite regress, because we must rely upon perceptual judgments to collect the data for an inductive generalization.) By contrast, Reiter's defaults often represent contingent generalizations. If we know that most birds can fly, then the inference from being a bird to flying may be adopted as a default. In the theory of defeasible reasoning propounded in this book, the latter inference is instead handled in terms of the statistical syllogism, as discussed in chapter 2. A second contrast between the present theory of defeasible reasoning and Reiter's approach is that the latter is semantical (proceeding in terms of an unspecified deductive-consequence relation), whereas the former is argument-theoretic.

It is easily proven that if we identify prima facie reasons with defaults, confine our attention to linear arguments, consider only cases in which there is no collective defeat,[2] and identify the deductive-consequence relation with deductive provability in OSCAR, then the set of warranted conclusions generated by the present theory will be the same as the unique extension generated by Reiter's default logic. In situations in which collective defeat occurs, the two theories yield completely different results, because default logic is credulous and the present theory is skeptical. Recall that in cases of collective defeat, a skeptical theory dictates withholding belief, whereas a credulous theory

---

[2] Collective defeat was introduced in chapter 2, and will be made precise in section 6 of this chapter.

dictates choosing one of the defeated conclusions at random and believing that. It was argued in chapter 2 that the standard defenses of credulous systems confuse epistemic and practical reasoning, and that credulous systems of epistemic reasoning are simply wrong. However, a skeptical version of default logic can be generated by requiring that default consequences be members of the intersection of all extensions, and this brings the two theories back into agreement on "simple" cases,[3] provided we consider only linear arguments. Once we allow suppositional reasoning, the two theories diverge again. For instance, if we consider a default theory with the default $P{:}Q/Q$ and the corresponding defeasible theory in which P is a prima facie reason for Q and there are no undercutting defeaters, then from the empty set of premises the present theory of defeasible reasoning will generate the warranted conclusion $(P \supset Q)$, but skeptical default logic will not.

Circumscription

Circumscription was developed by McCarthy [1980, 1984]. The basic idea is that if we know that most A's are B's, then in reasoning about A's, we should assume that the set of exceptional A's that are not B's is *minimal*. In other words, we assume that the only exceptions are those forced upon us by other things we know. McCarthy captures this by adding second-order axioms to the theory. He begins by symbolizing ⌜most A's are B's⌝ as

$$(\forall x)\{[A(x) \mathbin{\&} {\sim}ab(x)] \supset B(x)\}$$

where ⌜$ab(x)$⌝ symbolizes ⌜*x is abnormal (exceptional)*⌝. Letting $T(ab)$ be our initial theory about A's, B's, and abormality, the *circumscription of* T is the second-order principle:

$$T(ab) \mathbin{\&} (\forall X)\{[T(X) \mathbin{\&} (\forall x)(X(x) \supset ab(x))] \supset (\forall x)(ab(x) \supset X(x))\}.$$

This says that the theory holds for *ab*, and if the theory holds for any property X whose extension is a subset of the extension of *ab*, then the extension of *ab* is the same as the extension of X. In other words, the extension of *ab* is a minimal set sufficient to satisfy the theory T.

To illustrate, suppose our theory T is that most birds fly, and Tweety

---

[3] "Simple" cases are those that do not exhibit self-defeat, ancestor-defeat, or any of the other complexities discussed in sections 6–8.

is a bird. We can express this as:

$$bird(Tweety)$$
$$(\forall x)\{[bird(x) \ \& \ {\sim}ab(x)] \supset fly(x)\}$$

We want to infer, defeasibly, that Tweety flies. The circumscription of this theory tells us that the set of nonflying birds is a minimal set satisfying this theory. The minimal set satisfying this theory is the empty set, so from the circumscription we can infer that Tweety flies.

Although circumscription captures some of the same inferences as my theory of defeasible reasoning, it also has some peculiar consequences, which, to my mind, show that it is not an adequate theory. To adapt an example from Etherington, Kraus, and Perlis [1991], consider a lottery that is held once a week. Each week, one ticket is drawn. Most tickets are not drawn so the ticket drawn is abnormal, in the technical sense employed in circumscription. Circumscribing abnormality, there will be one minimal extension of *ab* for each ticket, corresponding to that ticket's being drawn. A conclusion follows from the circumscription only if it is true in each minimal extension, so no conclusion follows to the effect that any particular ticket will not be drawn. So far, circumscription yields the right answer.[4] But now modify the example. Each week, instead of drawing just one ticket, from one to five tickets are drawn, the number being determined randomly. The minimal extensions of *ab* are still those in which just one ticket is drawn, so it is true in each minimal extension that only one ticket is drawn. Thus, each week, circumscription allows us to conclude that only one ticket will be drawn. But this conclusion is unreasonable. Neither default logic nor the theory of defeasible reasoning adumbrated in this book has this consequence. The undesirable conclusion is directly traceable to circumscription's minimization of abnormality, so it seems to show that this is not, after all, a satisfactory way of capturing the structure of defeasible reasoning.

In the following sections, circumscription and (skeptical) default logic will be compared with the theory of defeasible reasoning in their application to a number of problem cases. I will argue that, in at least

---

[4] Etherington, Krause, and Perlis [1991] maintain that this is the wrong answer, and that it is reasonable to conclude of any given ticket that it will not be drawn. I disagree with them. See the discussion of the lottery paradox in chapter 2.

some cases, the former theories give unsatisfactory results.

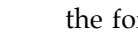
6. Computing Defeat Status

Justified beliefs are those supported by undefeated nodes of the inference graph. More precisely:

> A belief is justified to degree δ iff it is supported by some undefeated node of the inference graph whose strength is ≥ δ.

To complete this analysis, let us address the question of how the defeat status of a node of the inference graph is determined by its defeat relations to other nodes. It is initially plausible that there are just two ways a node can come to be defeated: (1) by its being defeated by some other node that is itself undefeated or (2) by its being inferred from a node that is defeated. Let us say that a node is *d-initial* iff neither it nor any of its inference ancestors are defeated by any nodes (that is, they are not the termini of any defeat links). D-initial nodes are guaranteed to be undefeated. Then we might try the following recursive definition:

(6.1)
1. D-initial nodes are undefeated.
2. If the immediate ancestors of a node η are undefeated and all nodes defeating η are defeated, then η is undefeated.
3. If η has a defeated immediate ancestor, or there is an undefeated node that defeats η, then η is defeated.

To illustrate, suppose we have the inference graph diagrammed in figure 2, where defeasible inferences are indicated by dashed arrows, deductive inferences by solid arrows, and defeat links by arrows of the form "  ". α,β,ε,ν,μ,κ, and λ are d-initial nodes, so they are undefeated. By (6.1.3), ζ, ξ, and χ are then defeated. By (6.1.2), because β is undefeated and ξ is defeated, γ and δ are then undefeated.

In simple cases, all standard theories of defeasible reasoning and nonmonotonic logic will yield results that are in agreement with principle (6.1), but as we will see, the different theories diverge on some complicated cases.

I take it that principle (6.1) is an initially plausible proposal for computing defeat status. However, the operation of this recursive definition is not as simple as it might at first appear. In figure 2,

principle (6.1) assigns "defeated" or "undefeated" to each node of the inference graph, but that will not always be the case. In particular, this will fail in cases of collective defeat, where we have a set of nodes, each of which is defeated by other members of the set and none of which is defeated by undefeated nodes outside the set. Consider the simple inference graph diagrammed in figure 3. In this case, α and ε are again d-initial nodes and hence undefeated. But neither β nor ζ will be assigned any status at all by principle (6.1), and then it follows that no status is assigned to any of γ,δ,ξ, or χ either.

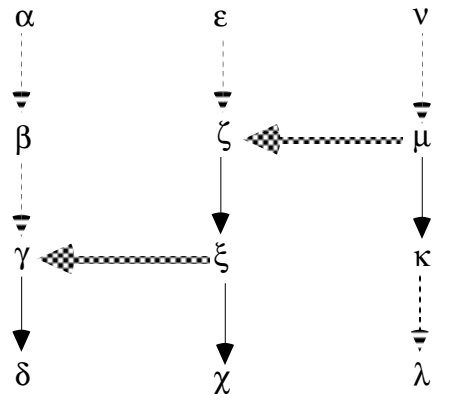


Figure 2. Inference graph

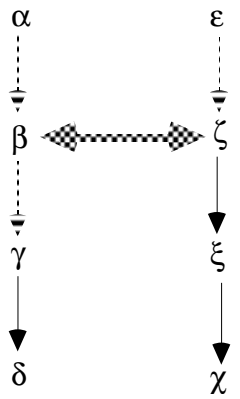


Figure 3. Collective defeat

Collective defeat was illustrated in chapter 2 by the lottery paradox. Suppose you hold one ticket in a fair lottery consisting of 1 million

tickets, and suppose it is known that one and only one ticket will win. Observing that the probability is only .000001 of a ticket's being drawn given that it is a ticket in the lottery, it seems reasonable to accept the conclusion that your ticket will not win. The prima facie reason involved in this reasoning is the statistical syllogism. But by the same reasoning, it will be reasonable to believe, for each ticket, that it will not win. However, these conclusions conflict jointly with something else we are justified in believing: that some ticket will win. We cannot be justified in believing each member of an explicitly contradictory set of propositions, and we have no way to choose between them, so it follows intuitively that we are not justified in believing of any ticket that it will not win.[5] This is captured formally by the principle of collective defeat, which tells us that our prima facie reasons collectively defeat one another:

> ***The Principle of Collective Defeat***
> If X is a set of nodes of the inference graph, each member of X is defeated by another member of X, and no member of X is defeated by an undefeated node that is not a member of X, then every node in X is defeated.

Collectively defeated inferences are defeated, in the sense that it is unreasonable to accept their conclusions.[6] But principle (6.1) does not rule them defeated. This may be less of a problem for principle (6.1) than it seems. We can regard the assignment of defeat statuses in figure 3 as correct, provided we go on to say that β and ζ should be assigned a third status distinct from both "defeated" and "undefeated". The need for a third defeat status is best illustrated by contrasting figure 2 with figure 4. In figure 2, ζ and hence ξ are defeated, and ξ thereby loses the ability to render γ defeated. In figure 4, both ζ and μ are defeated (it would not be reasonable to accept their conclusions), but ζ retains the ability to render β defeated, because it would not be

---

[5] Kyburg [1970] draws a different conclusion: we can be justified in holding inconsistent sets of beliefs, and it is not automatically reasonable to adopt the conjunction of beliefs one justifiably holds (i.e., adjunction fails). This was discussed in section 2.

[6] Not everyone accepts this diagnosis. For a defense of it, see my discussion of the lottery paradox in chapter 2.

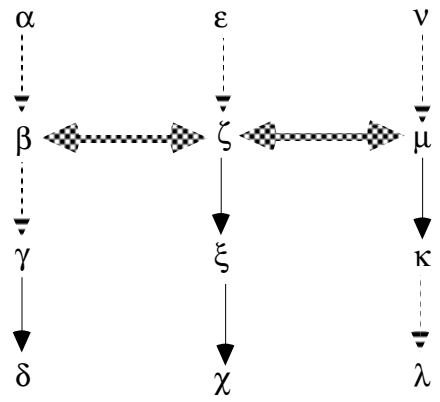


Figure 4. The propagation of provisional defeat

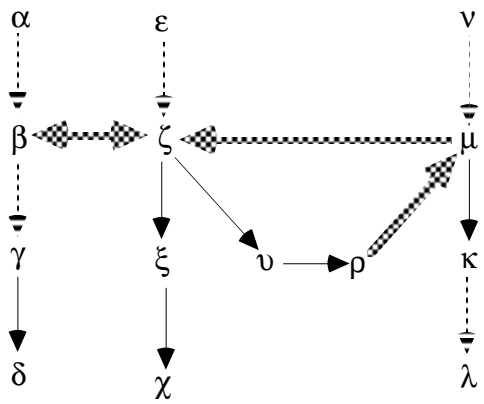


Figure 5. The extended propagation of provisional defeat

reasonable to accept the conclusion of β either. This is an unavoidable

consequence of the symmetry of the inference graph. The relationship between β and ζ is precisely the same as that between ζ and μ. We must regard both as cases of collective defeat. The order in which the arguments are produced, or the nodes considered by the recursion, cannot affect their defeat status.

We can handle this by distinguishing between two kinds of defeat: *outright defeat* and *provisional defeat*. If a node undergoes outright defeat, it loses the ability to affect other nodes, but if a node undergoes provisional defeat, it can still render other nodes provisionally defeated. Provisionally defeated nodes are still "infectious". Provisional defeat can propagate, in two ways. First, as illustrated by figure 4, if a provisionally defeated node defeats a node that would not otherwise be defeated, this can render the latter node provisionally defeated. Second, if a node is inferred from a provisionally defeated node and its other immediate ancestors are undefeated, then that node may be provisionally defeated as well. That is, a node inferred from a provisionally defeated node is defeated but may still be infectious. This is illustrated by making structures like figure 4 more complicated so that the collective defeat of ζ and μ involves extended reasoning, as in figure 5. Here, υ and ρ are inferred from ζ, so they are defeated, but they must remain infectious in order to defeat μ and thus generate the provisional defeat of ζ and μ. If υ and ρ were defeated outright rather than provisionally, then μ would be undefeated, which would render ζ defeated outright, but that is intuitively wrong.

Outright defeat and provisional defeat are both defeat, in the sense that it is not reasonable to accept the conclusion of a node with either status. But the two defeat statuses are importantly different in that a node is rendered impotent if it is defeated outright, but if it is only provisionally defeated, it retains the ability to defeat other nodes.

The examples considered thus far can be handled by adding a fourth clause to principle (6.1):

(6.2)
1. D-initial nodes are undefeated.
2. If the immediate ancestors of a node η are undefeated and all nodes defeating η are defeated outright, then η is undefeated.
3. If η has an immediate ancestor that is defeated outright, or there is an undefeated node that defeats η, then η is defeated outright.
4. Otherwise, η is provisionally defeated.

The automatic consequence is that otherwise undefeated nodes inferred from provisionally defeated nodes are provisionally defeated, and otherwise undefeated nodes defeated by provisionally defeated nodes

are provisionally defeated. Principle (6.2) is equivalent to the analysis of defeat I have given elsewhere [1979, 1986, 1987]. However, I now believe that this account is inadequate, for the reasons that will be explored next.

7. Self-defeating Arguments

The inadequacy of principle (6.2) can be illustrated by a wide variety of examples. The simplest is the following. Suppose P is a prima facie reason for R, Q is a prima facie reason for $\sim R$, S is a prima facie reason for T, and we are given P,Q, and S. Then we can do the reasoning encoded in the inference graph diagrammed in figure 6. The nodes (R) and $(\sim R)$ collectively defeat one another, but (T) should be independent of either and undefeated. The difficulty is that we can extend the inference graph as in figure 7. Here I have used a standard strategy for deriving an arbitrary conclusion from a contradiction. Now the problem is that $(\sim T)$ rebuts (T). According to principle (6.2), $(\sim R)$, and hence $(\sim T)$, are provisionally defeated, but then it follows that (T) is also provisionally defeated. The latter must be wrong. There are no constraints on T, so it would have the consequence that all conclusions are defeated. This example shows that nodes inferred from provisionally defeated nodes are not always provisionally defeated. In figure 7, $(\sim T)$ must be defeated outright. There is no way to get this result from principle (6.2). My diagnosis of the difficulty is that the argument supporting $(\sim T)$ is "internally defective". It is *self-defeating* in the sense that some of its steps are defeaters for others. By principle (6.2), this means that those inferences enter into collective defeat with one another, and hence $(\sim T)$ is provisionally defeated, but my suggestion is that this should be regarded as a more serious defect—one that leaves $(\sim T)$ defeated outright and hence unable to defeat other inferences. Taking the *inclusive inference ancestors* of a node to be its inference ancestors together with itself, let us define:

> A node η is *self-defeating* iff some of its inclusive inference ancestors defeat others.

Principle (6.2) should be modified so that self-defeating nodes are defeated outright rather than just provisionally.

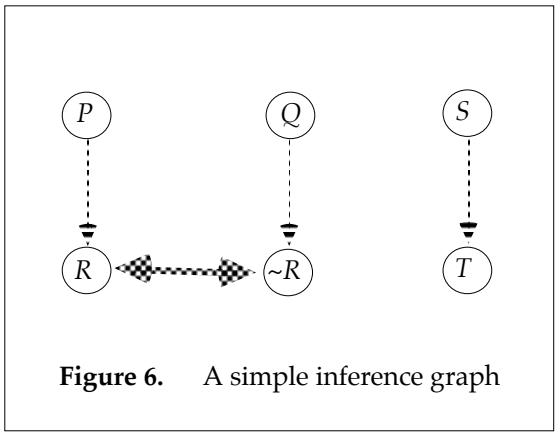


Figure 6. A simple inference graph

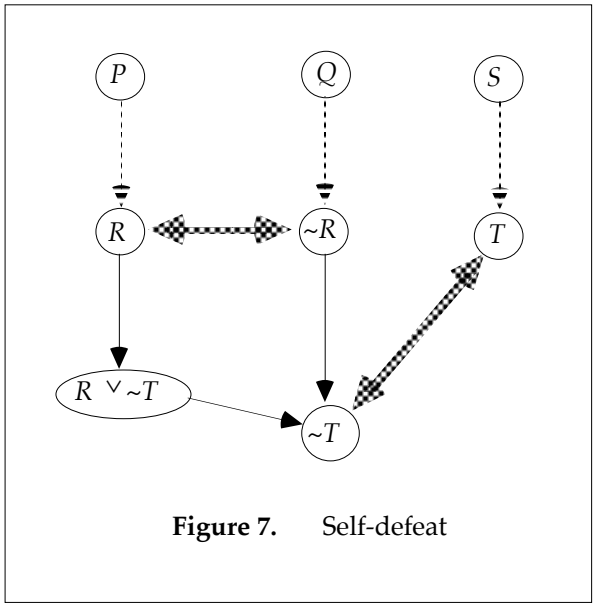


Figure 7. Self-defeat

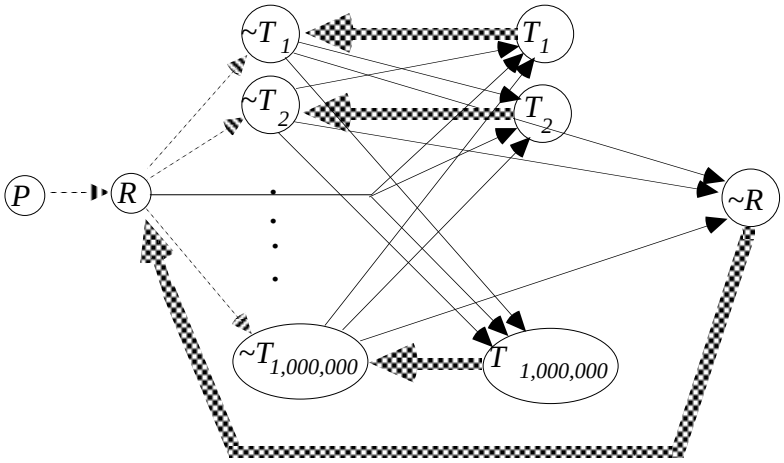


Figure 8. Lottery Paradox Paradox

It is noteworthy that neither (skeptical) default logic nor circumscription has any difficulty with the inference graph of figure 7. In default logic, there is one minimal extension containing *R* and another containing *~R* but no minimal extension containing both and so none containing *~T*. Similarly, in circumscribing abnormality, either the inference to *R* or the inference to *~R* will be blocked by abnormality, and in either case the inference to *~T* will be blocked.

Circumscription does not fare so well when we turn to a second example of self-defeat that has a somewhat different structure. This concerns what appears to be a paradox of defeasible reasoning and involves the lottery paradox again. The lottery paradox is generated by supposing that a proposition *R* describing the lottery (it is a fair lottery, has 1 million tickets, and so on) is justified. Given that *R* is justified, we get collective defeat for the proposition that any given ticket will not be drawn. But principle (6.2) makes it problematic how *R* can be justified. Normally we will have only a defeasible reason for believing *R*. For instance, we may be told that it is true or read it in a newspaper. 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 generate an argument supporting *~R* by noting that the $\sim T_i$ jointly entail *~R*, because if none of the tickets is drawn, the lottery is not fair. This is diagrammed in figure 8. The difficulty is now that (*~R*) rebuts (*R*). Thus by principle (6.2), these nodes defeat one another,

with the result that neither is defeated outright. In other words, the inference to *R* is provisionally defeated. Again, this result is intuitively wrong. Oviously, if we consider examples of real lotteries (e.g., this week's New York State Lottery), it is possible to become justified in believing *R* on the basis described. I propose once more that the solution to this problem lies in noting that the node $(\sim R)$ is self-defeating.

Default logic gets the example of figure 8 right, but circumscription gets it wrong. In circumscribing abnormality, all we can conclude is that one of the defeasible inferences is blocked by abnormality, but it could be the inference to *R*, so circumscription does not allow us to infer *R*.

On the argument-based approach, the difficulties diagrammed in figures 7 and 8 can be avoided by ruling that self-defeating nodes are defeated outright—not just provisionally. Because they are defeated outright, they cannot enter into collective defeat with other nodes, and so the nodes $(\sim R)$ and $(\sim T)$ in the preceding two examples are defeated outright, as they should be. This can be accomplished by revising principle (6.2) as follows:

(7.1)
1. D-initial nodes are undefeated.
2. Self-defeating nodes are defeated outright.
3. If η is not self-defeating, its immediate ancestors are undefeated, and all nodes defeating η are defeated outright, then η is undefeated.
4. If η has an immediate ancestor that is defeated outright, or there is an undefeated node that defeats η, then η is defeated outright.
5. Otherwise, η is provisionally defeated.

This is equivalent to one of the preliminary proposals made in my [1990]. However, it is still inadequate.

An interesting problem arises when the last step of an argument constitutes an undercutting defeater for an earlier step. Consider the inference graph diagrammed in figure 9. The node $(P \otimes Q)$ is self-defeating, because it defeats one of its own ancestors. Thus by principle (7.1), it is defeated outright. It then follows from principle (7.1) that the remaining nodes are undefeated. But this is most peculiar, because $(P \otimes Q)$ is a deductive consequence of (R). If a node is undefeated, its deductive consequences should also be undefeated. Conversely, if

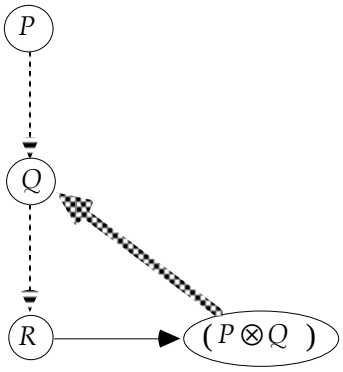


Figure 9. No nearest defeasible ancestor is defeated

a node is inferred deductively from a set of nodes (its *nearest defeasible ancestors*), then if the node is defeated, at least one of its nearest defeasible ancestors should also be defeated. It follows that at least (R) should be defeated. What about (Q)? Intuitions are unclear in such an abstract example, so let us turn to a concrete example.

Suppose we know (*i*) that people generally tell the truth, (*ii*) that Robert says that the elephant beside him looks pink, and (*iii*) that Robert becomes unreliable in the presence of pink elephants. ⌜x looks pink⌝ is a prima facie reason for ⌜x is pink⌝. Then Robert's statement gives us a prima facie reason for thinking that the elephant *does* look pink, which gives us a reason for thinking that it *is* pink, which, when combined with Robert's unreliability in the presence of pink elephants, gives us a defeater for our reason for thinking that the elephant looks pink. These relations can be diagrammed as in figure 10. Node 7 is self-defeating, so one of its nearest defeasible ancestors ought to be defeated. These are nodes 5 and 6. Of these, it seems clear that node 6 should be defeated *by* having node 4 defeated. That is, in this example, it would not be reasonable to accept the conclusion that the elephant beside Robert looks pink. This strongly suggests that we should similarly regard (Q) as defeated in figure 9. Neither of these conclusions is forthcoming from principle (7.1). In earlier publications I tried to resolve these problems by generalizing the notion of self-defeat, but I no longer believe that those attempts were successful.

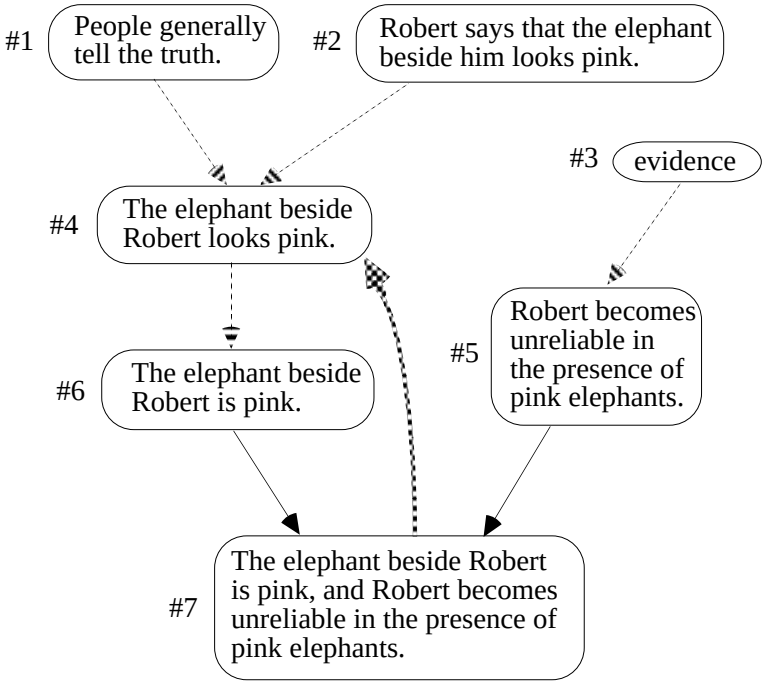


Figure 10. Argument with a self-defeating conclusion

It turns out that circumscription gives the right result in figures 9 and 10. In figure 9, circumscribing abnormality has the consequence that either the inference to Q or the inference to R is blocked, and hence Q does not follow from the circumscription. On the other hand, default logic gives an outlandish result in figure 9. It turns out that there are *no* extensions in this case, and hence either nothing is justified (including the given premise P) or everything is justified, depending upon how we handle this case. This seems to be a fairly clear counter-example to default logic.

8. A New Approach

Default logic and circumscription handle some of the problem cases correctly and some incorrectly. The cases in which they fail tend to be

ones in which they are not sufficiently sensitive to the structure of the arguments. For example, in figure 8, circumscription gets self-defeat wrong, and in figure 9, default logic gets self-defeat wrong. This suggests that the argument-based approach should be superior, but as we have seen, the formulations of the argument-based approach that are contained in principles (6.2) and (7.1) fail to deal adequately with at least one of the examples that default logic and circumscription get right. The attempts to salvage the argument-based approach by building in restrictions become increasingly ad hoc as the examples become more complex. I think it is time to abandon the search for such restrictions and look for another way of handling the problems. Here, I think that the argument-based approach has a lesson to learn from default logic and circumscription. Consider the first example of collective defeat—figure 7. Default logic and circumscription get this example right, but principle (6.2) gets it wrong, necessitating the explicit appeal to self-defeat in principle (7.1). It is illuminating to consider why default logic and circumscription have no difficulty with this example. This is because they take account of the relationship between the provisionally defeated conclusions R and $\sim R$ instead of just throwing them all into an unstructured pot of provisionally defeated conclusions. This allows us to observe that when R is "acceptable", $\sim R$ is not, and hence there are no circumstances under which $\sim T$ is "acceptable". Principle (6.2), on the other hand, washes these relationships out, assigning a blanket status of "provisionally defeated" to all provisionally defeated propositions.

The conclusion I want to draw is that the argument-based approach gets things partly right, and default logic and circumscription get things partly right. What is needed is a single theory that combines the insights of both. In order to take account of the structure of arguments, this will have to be an argument-based theory, but in assessing defeat statuses, it must take account of the interconnections between nodes and not just look at the defeat statuses of the nodes that are inference ancestors or defeaters of a given node. There is a way of taking account of such interconnections while remaining within the spirit of principles (6.1) and (6.2). Let us define a *status assignment* to be an assignment of defeat status that is consistent with the rules of principle (6.1). When nodes are either undefeated or defeated outright, then every status assignment will accord them that status, but when nodes are provisionally defeated, some status assignments will assign the status "defeated" and others the status "undefeated". Links between nodes will be reflected in the fact that, for example, every status assignment making one undefeated may make another defeated. This is made precise as follows:

An assignment σ of "defeated" and "undefeated" to the nodes of an inference graph is a *status assignment* iff:

1. σ assigns "undefeated" to all d-initial nodes;
2. σ assigns "undefeated" to a node α iff σ assigns "undefeated" to all the immediate ancestors of α and all nodes defeating α are assigned "defeated"; and
3. σ assigns "defeated" to α iff either α has a immediate ancestor that is assigned "defeated", or there is a node β that defeats α and is assigned "undefeated".

The proposal is then:

(8.1) A node is undefeated iff every status assignment assigns "undefeated" to it; otherwise it is defeated. Of the defeated nodes, a node is defeated outright iff no status assignment assigns "undefeated" to it; otherwise, it is provisionally defeated.

This simple proposal deals adequately with all but one of the examples we have considered. In figure 7, there is one status assignment assigning "defeated" to (R) and "undefeated" to $(\sim R)$, and another status assignment assigning the opposite statuses. On both assignments, $(\sim T)$ is assigned "defeated", so by principle (8.1), $(\sim T)$ is defeated outright. Figure 8 is analogous. In figure 8, for each i there is a status assignment assigning "defeated" to $(\sim T_i)$ but assigning "undefeated" to (R) and all the other $(\sim T_j)$'s. Every such status assignment assigns "defeated" to $(\sim R)$. Thus by principle (8.1), (R) is undefeated and $(\sim R)$ is defeated outright, while all of the $(\sim T_i)$'s are provisionally defeated.

Of the above examples, principle (8.1) is able to handle all but that of figure 9. In figure 9, something unexpected happens. Any status assignment assigning "undefeated" to (Q) will also assign "undefeated" to (R) and to $((P \otimes Q))$, but then it must instead assign "defeated" to (Q). Thus no status assignment can assign "undefeated" to (Q). However, no status assignment can assign "defeated" to (Q) either, because then it would have to assign "defeated" to (R) and $((P \otimes Q))$ as well, from which it follows that it must instead assign "undefeated" to (Q). This shows that no status assignments are possible for the inference graph of figure 9. We can construct other examples of this same

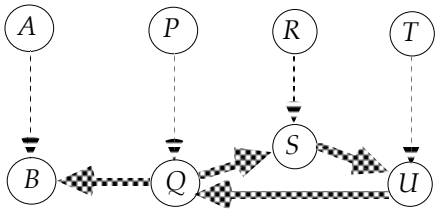


Figure 11. A three-membered defeat cycle.

phenomenon. The simplest involve odd-length defeat cycles. Consider the inference graph diagrammed in figure 11. For example, we might let P be "Jones says that Smith is unreliable", Q be "Smith is unreliable", R be "Smith says that Robertson is unreliable", S be "Robertson is unreliable", T be "Robertson says that Jones is unreliable", U be "Jones is unreliable", and A be "Smith says that it is raining" and B be "It is raining". Intuitively, $\langle Q \rangle$, $\langle S \rangle$, and $\langle U \rangle$ ought to defeat one another collectively, and then because $\langle Q \rangle$ is provisionally defeated, $\langle B \rangle$ should be provisionally defeated. That is precisely the result we get if we expand the defeat cycle to four nodes. However, in the inference graph containing the three-membered defeat cycle, there is no way to assign defeat statuses consistent with principle (6.1). For example, if $\langle Q \rangle$ is assigned "undefeated", $\langle S \rangle$ must be assigned "defeated", and then $\langle U \rangle$ must be assigned "undefeated", with the result that $\langle Q \rangle$ must be assigned "defeated"—a contradiction. Every other way of trying to assign defeat statuses yields a similar contradiction. Consequently, there is no status assignment for the inference graph in figure 11. But surely, it should make no difference that the defeat cycle is of odd length rather than even length. We should get the same result in either case.

This difficulty can be rectified by allowing status assignments to be partial assignments. They can leave gaps, but only when there is no consistent way to avoid that. Accordingly, let us revise the earlier definition as follows:

> An assignment σ of "defeated" and "undefeated" to a subset of the nodes of an inference graph is a *partial status assignment* iff:
> 1. σ assigns "undefeated" to all d-initial nodes;
> 2. σ assigns "undefeated" to a node α iff σ assigns "undefeated" to all the immediate ancestors of α and all nodes de-

feating α are assigned "defeated"; and

3. σ assigns "defeated" to a node α iff either α has a immediate ancestor that is assigned "defeated", or there is a node β that defeats α and is assigned "undefeated".

Status assignments are then maximal partial status assignments:

> σ is a *status assignment* iff σ is a partial status assignment and σ is not properly contained in any other partial status assignment.

With this modification, principle (8.1) handles the examples of figures 9 and 11 properly. In figure 9, nodes (Q), (R), and ((P ⊗ Q)) turn out to be (provisionally) defeated, and in figure 11, nodes (B), (Q), (Q), and (U) are (provisionally) defeated. This is my final proposal for the analysis of defeat for nodes of the inference graph.[7]

This analysis entails that defeat statuses satisfy a number of intuitively desirable conditions:

(8.2) A node α is undefeated iff all immediate ancestors of α are undefeated and all nodes defeating α are defeated.

(8.3) If some immediate ancestor of α is defeated outright then α is defeated outright.

(8.4) If some immediate ancestor of α is provisionally defeated, then α is either provisionally defeated or defeated outright.

(8.5) If some node defeating α is undefeated, then α is defeated outright.

(8.6) If α is self-defeating then α is defeated outright.

9. The Paradox of the Preface

Much of my work on the analysis of defeat has been driven by an

---

[7] In a number of earlier publications [1979, 1986, 1987, 1990, 1990c, 1991, and 1992], I proposed that defeat could analyzed as defeat among *arguments* rather than inference nodes, and I proposed an analysis of that relation in terms of "levels of arguments". I now believe that obscured the proper treatment of self-defeat and ancestor defeat. I see no way to recast the present analysis in terms of a defeat relation between arguments (as opposed to nodes, which are argument steps rather than complete arguments).

attempt to deal adequately with the lottery paradox and the paradox of the preface. The difficulty is that these two paradoxes seem superficially to have the same form, and yet they require different resolutions. I discussed the lottery paradox above, maintaining that it can be regarded as a straightforward case of collective defeat. Contrast that with the paradox of the preface [Makinson 1965], which can be presented as follows:

> There once was a man who wrote a book. He was very careful in his reasoning, and was confident of each claim that he made. With some display of pride, he showed the book to a friend (who happened to be a probability theorist). He was dismayed when the friend observed that any book that long and that interesting was almost certain to contain at least one falsehood. Thus it was not reasonable to believe that all of the claims made in the book were true. If it were reasonable to believe each claim then it would be reasonable to believe that the book contained no falsehoods, so it could not be reasonable to believe each claim. Furthermore, because there was no way to pick out some of the claims as being more problematic than others, there could be no reasonable way of withholding assent to some but not others. "Therefore," concluded his friend, "you are not justified in believing anything you asserted in the book." [Pollock 1991]

This is the paradox of the preface (so named because in the original version the author confesses in the preface that his book probably contains a falsehood). This paradox is made particularly difficult by its similarity to the lottery paradox. In both paradoxes, we have a set Γ of propositions, each of which is supported by a defeasible argument, and a reason for thinking that not all of the members of Γ are true. But in the lottery paradox we want to conclude that the members of Γ undergo collective defeat, and hence we are not justified in believing them, whereas in the paradox of the preface we want to insist that we are justified in believing the members of Γ. How can the difference be explained?

There is, perhaps, some temptation to acquiesce in the reasoning involved in the paradox of the preface and conclude that we are not justified in believing any of the claims in the book after all. That would surely be paradoxical, because a great deal of what we believe about the world is based upon books and other sources subject to the same argument. For instance, why do I believe that Alaska exists? I have never been there. I believe it only because I have read about it. If the reasoning behind the paradox of the preface were correct, I

would not be justified in believing that Alaska exists. That cannot be right.

The paradox of the preface may seem like an esoteric paradox of little more than theoretical interest. However, the *form* of the paradox of the preface is of fundamental importance to defeasible reasoning. That form recurs throughout defeasible reasoning, with the result that if that form of argument were not defeated, virtually all beliefs based upon defeasible reasoning would be unjustified. This arises from the fact that we are typically able to set at least rough upper bounds on the reliability of our prima facie reasons. For example, color vision gives us prima facie reasons for judging the colors of objects around us. Color vision is pretty reliable, but surely it is not more than 99.9% reliable. Given that assumption, it follows that the probability that out of 10,000 randomly selected color judgments, at least one is incorrect, is 99.99%. By the statistical syllogism, that gives us a prima facie reason for thinking that at least one of them is false. By reasoning analogous to the paradox of the preface, it seems that none of those 10,000 judgments can be justified. And because every color judgment is a member of some such set of 10,000, it follows that all color judgments are unjustified. The same reasoning would serve to defeat any defeasible reasoning based upon a prima facie reason for which we can set at least a rough upper bound of reliability. Thus it becomes imperative to resolve the paradox of the preface.

The paradox of the preface can be resolved by appealing to the analysis of defeat proposed above.[8] The paradox has the following form. We begin with a set $\Gamma = \{p_1,...,p_N\}$ of propositions, where Γ has some property B (being the propositions asserted in a book of a certain sort, or being a set propositions supported by arguments employing a certain prima facie reason). We suppose we know that the probability of a member of such a set being true is high, but we also know that it is at least as probable that such a set of propositions contains at least one false member. Letting T be the property of being true, we can express these probabilities as:

$$\text{prob}(Tz \mid z \in X \mathbin{\&} B(X)) = r$$
$$\text{prob}((\exists z)(z \in X \mathbin{\&} \sim Tz) \mid B(X)) \geq r.$$

The latter high probability, combined with the premise $B(\Gamma)$, gives us

[8] This corrects my earlier discussion [1990, 1991].

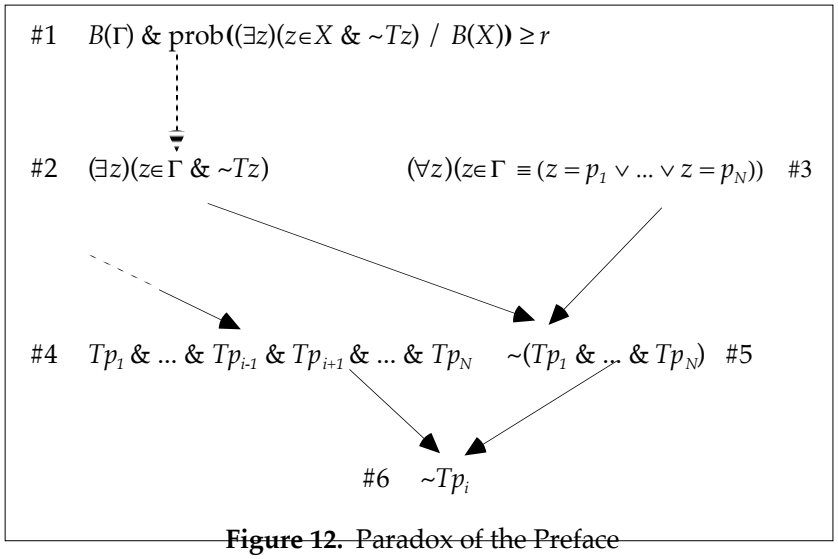


Figure 12. Paradox of the Preface

a defeasible reason for $(\exists z)(z\in\Gamma\ \&\ {\sim}Tz)$. This, in turn, generates collective defeat for all the arguments supporting the members of Γ. The collective defeat is generated by constructing the argument scheme diagrammed in figure 12 for each ${\sim}Tp_i$.

A resolution of the paradox of the preface must consist of a demonstration that node 6 is defeated outright. A subproperty defeater for the reasoning from node 1 to node 2 arises from establishing anything of the following form (for any property C):

$$C(\Gamma)\ \&\ \text{prob}((\exists z)(z\in X\ \&\ {\sim}Tz)\ /\ B(X)\ \&\ C(X)) < r.^{9}$$

I have shown [1990, 251] that

$$\text{prob}((\exists z)(z\in X\ \&\ {\sim}Tz)\ /\ B(X)\ \&\ X = \{x_1,...,x_N\}\ \&\ x_1,...,x_N \text{ are distinct}^{10}\ \&\ Tx_1\ \&\ ...\ \&\ Tx_{i-1}\ \&\ Tx_{i+1}\ \&\ ...\ \&\ Tx_N)$$

$$= \text{prob}({\sim}Tx_i\ /\ B(X)\ \&\ X = \{x_1,...,x_N\}\ \&\ x_1,...,x_N \text{ are distinct}\ \&\ Tx_1\ \&\ ...\ \&\ Tx_{i-1}\ \&\ Tx_{i+1}\ \&\ ...\ \&\ Tx_N).$$

Now we come to the point at which the paradox of the preface differs

---

[9] See section 6 of chapter 2 for more details on subproperty defeaters.

[10] "$x_1,...,x_n$ are distinct" means "$x_1,...,x_n$ are n different objects".

from the lottery paradox. In the lottery paradox, knowing that none of the other tickets has been drawn makes it likely that the remaining ticket is drawn. By contrast, knowing that none of the other members of Γ is false does not make it likely that the remaining member of Γ is false. In other words,

$$\begin{aligned}&\text{prob}(\sim Tx_i \,/\, B(X) \;\&\; X = \{x_1,\ldots,x_N\} \;\&\; x_1,\ldots,x_N \text{ are distinct } \&\\ &\qquad Tx_1 \;\&\; \ldots \;\&\; Tx_{i-1} \;\&\; Tx_{i+1} \;\&\; \ldots \;\&\; Tx_N)\end{aligned}$$

$$\leq \text{prob}(\sim Tx_i \,/\, B(X) \;\&\; X = \{x_1,\ldots,x_N\} \;\&\; x_1,\ldots,x_N \text{ are distinct}).$$

Equivalently, the different claims in Γ are not negatively relevant to one another. For example, the 10,000 color judgments were assumed to be independent of one another, so these two probabilities are equal in that case. In the case of the book, the various claims would normally be taken to support one another, if anything, and so be positively relevant rather than negatively relevant. There is no reason to believe that the condition $\ulcorner X = \{x_1,\ldots,x_N\} \;\&\; x_1,\ldots,x_N \text{ are distinct}\urcorner$ alters the probability, so it is reasonable to believe that the last-mentioned probability is just $1–r$, which, of course, is much smaller than r.[11] Thus we have

$$\begin{aligned}&\text{prob}((\exists z)(z \in X \;\&\; \sim Tz) \;/\; B(X) \;\&\; X = \{x_1,\ldots,x_N\} \;\&\; x_1,\ldots,x_N \text{ are}\\ &\qquad \text{distinct } \&\; Tx_1 \;\&\; \ldots \;\&\; Tx_{i-1} \;\&\; Tx_{i+1} \;\&\; \ldots \;\&\; Tx_N) < r.\end{aligned}$$

Accordingly, the conjunction

$$\begin{aligned}&\text{prob}((\exists z)(z \in X \;\&\; \sim Tz) \;/\; B(X) \;\&\; X = \{x_1,\ldots,x_N\} \;\&\; x_1,\ldots,x_N \text{ are}\\ &\qquad \text{distinct } \&\; Tx_1 \;\&\; \ldots \;\&\; Tx_{i-1} \;\&\; Tx_{i+1} \;\&\; \ldots \;\&\; Tx_N) < r\\ &\&\; p_1,\ldots,p_N \text{ are distinct}\end{aligned}$$

is warranted. Combining this with nodes 3 and 4 generates a subproperty defeater for the defeasible inference from node 1 to node 2, as diagrammed in figure 13. Consequently, node 8 defeats node 2.

---

[11] This inference proceeds by non-classical direct inference. See section 6 of chapter 2.

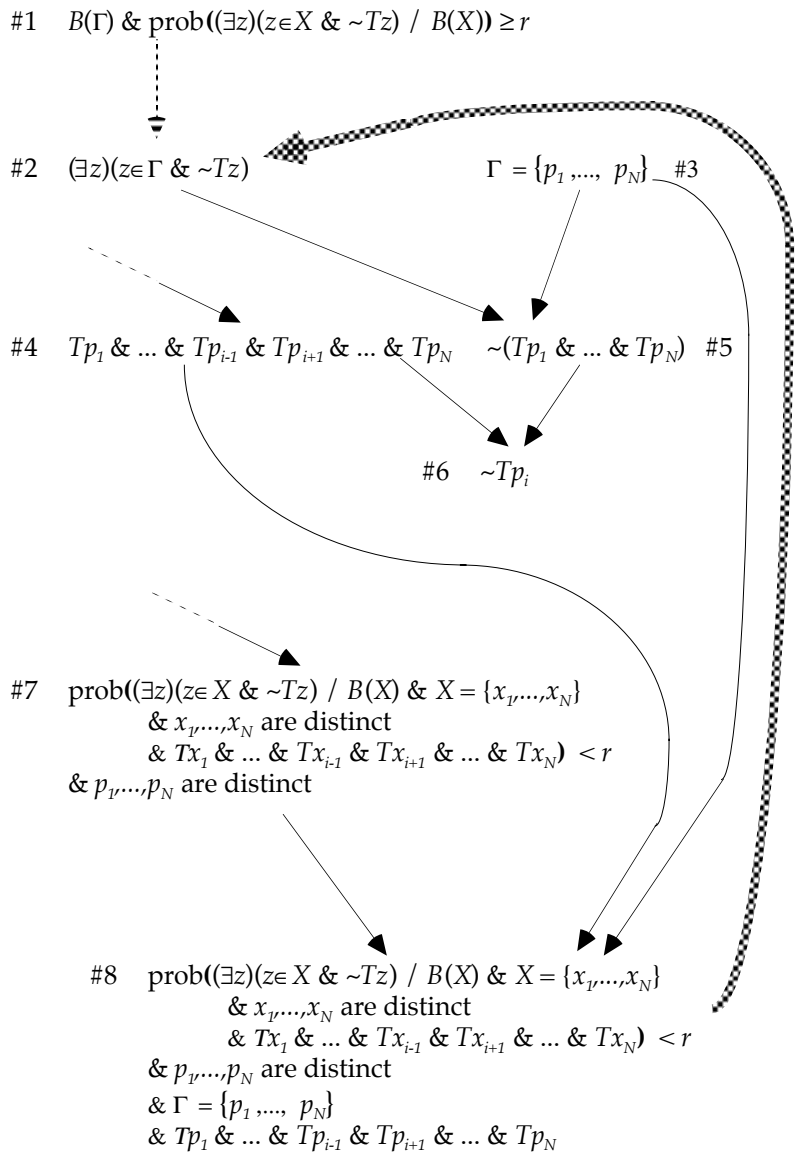


Figure 13. Resolution of the paradox of the preface

In computing defeat statuses, it is difficult to see the structure of this problem because there isn't room on a page to draw the entire inference graph. Figure 13 is only a partial diagram, because it does not take account of how the different Tp_i's are related to one another. The structure can be made clearer by considering a simpler problem having the same structure but with only three propositions playing the role of Tp_i's rather than a large number of them. Consider the inference graph diagrammed in figure 14. The pie-shaped regions are drawn in to emphasize the symmetry. The nodes supporting P_1, P_2, P_3, S, T and R are d-initial and hence undefeated. In evaluating the other nodes, note first that there is a status assignment assigning "undefeated" to the nodes supporting Q_1, Q, and Q_3. This assigns "undefeated" to the nodes supporting S_1, S_2, and S_3, and "defeated" to the nodes supporting $\sim Q_1$, $\sim Q_2$, $\sim Q_3$, and $\sim(Q_1 \,\&\, Q_2 \,\&\, Q_3)$. On the other hand, there can be no status assignment assigning "defeated" to the nodes supporting two different Q_i's, say Q_1 and Q_2, because if the latter were defeated, all nodes defeating the former would be defeated, and vice versa. (Note that this would still be true if there were more than three Q_i's.) Suppose instead that a status assignment assigns "defeated" to just one Q_i, and "undefeated" to the others. Then the node supporting S_i must be assigned "undefeated", and so the node supporting $\sim(Q_1 \,\&\, Q_2 \,\&\, Q_3)$ must be assigned "defeated". The result is that the node supporting $\sim Q_i$ must be assigned "defeated". That is the only node defeating that supporting Q, so the latter must be assigned "undefeated" after all. Hence there can be no node assigning "defeated" to a single Q_i. The result is that there is only one status assignment, and it assigns "undefeated" to the nodes supporting Q_1, Q_2, Q_3, S_1, S_2, and S_3, and "defeated" to the nodes supporting $\sim Q_1$, $\sim Q_2$, $\sim Q_3$, and $\sim(Q_1 \,\&\, Q_2 \,\&\, Q_3)$. Consequently, the former nodes are undefeated, and the latter are defeated outright.

This computation of defeat status can be applied to the paradox of the preface by taking the Q_i's to correspond to the nodes supporting each Tp_i. The nodes supporting the S_i's correspond to node 8 in figure 13. The nodes supporting the $\sim Q_i$'s correspond to note 6, the node supporting $\sim(Q_1 \,\&\, Q_2 \,\&\, Q_3)$ corresponds to nodes 2 and 5, and node supporting T corresponds to node 1, the node supporting R corresponds to node 7, and the nodes supporting the conjunctions of the form $(Q_i \,\&\, Q_j)$ correspond to node 4. Then a diagnosis analogous to that given for figure 14 yields the result that node 2, and hence node 6, are

both defeated outright, while the nodes supporting the Tp_i's are undefeated. It follows that the conjunction $(Tp_1 \,\&\, ... \,\&\, Tp_N)$ is justified. In

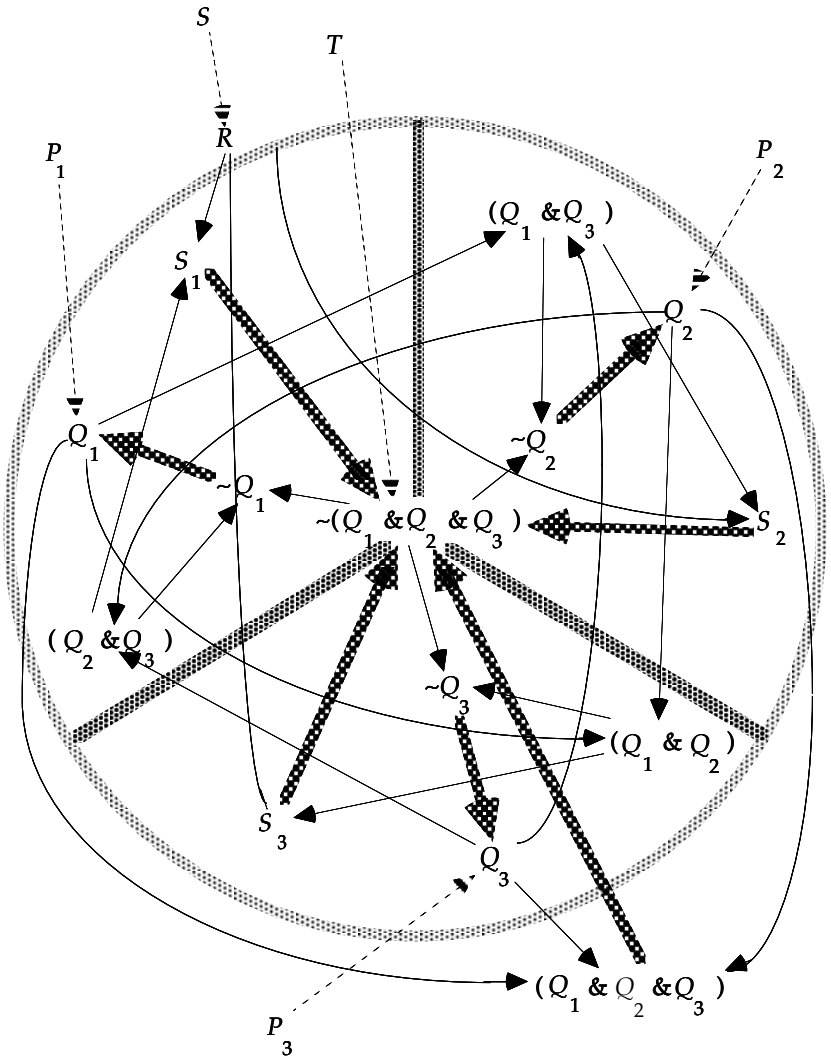


Figure 14. Structure of the paradox of the preface

other words, in the paradox of the preface, we are justified in believing that all the propositions asserted in that particular book are true, despite the fact that this is a book of a general type which usually contains some falsehoods.

If this still seems paradoxical, it is probably because one is overlooking the fact that "Books of this general sort usually contain falsehoods" formulates an *indefinite probability*, but "This book probably contains a falsehood" expresses a *definite* (single case) probability. The relationship between indefinite probabilities and definite probabilities is one of direct inference, which is a defeasible relation. In this case it is defeated by the fact that every proposition in the book is warranted, and hence the probability of *this* book's containing a falsehood is zero. (Warranted propositions automatically have probability one.) For more on direct inference, see section 6 of chapter 2.

This rather complex analysis shows that the difference between the paradox of the preface and the lottery paradox lies in the fact that the truth of the other propositions asserted in the book is not negatively relevant to the truth of the remaining proposition, but the other tickets in the lottery not being drawn *is* negatively relevant to the remaining ticket's not being drawn. This difference makes it reasonable to believe all the propositions asserted in the book but unreasonable to believe that none of the tickets will be drawn. This is also what makes it reasonable for us to believe our eyes when we make judgments about our surroundings. It is the analysis of defeat embodied in principle (8.1) that enables us to draw these congenial conclusions.

10. Justification and Warrant

A conclusion is justified if it is supported by an undefeated node of the inference graph. Reasoning starts with the set *input* and then produces new nodes to be added to the inference graph. Given a fixed *input*, let $\alpha_1,...,\alpha_i,...$ be the nodes that would be constructed by the reasoner (if it had unlimited resources and could continue reasoning forever), in the order in which they would be produced, and let G_i be the set of the first i nodes, $\{\alpha_1,...,\alpha_i\}$. Conclusions justified at the ith stage of reasoning are just those supported by nodes in G_i that are undefeated relative to G_i. Recall that the strength of a node is the minimum of the reason strengths of the reasons used in its inference ancestors and the strengths of the members of *input* that are inference ancestors of it. Epistemic justification is then made precise as follows:

> A sequent S is *justified to degree* δ *at stage* i iff there is a node α of strength greater than or equal to δ such that (1) $\alpha \in \mathbf{G}_i$, (2) α is undefeated relative to $\mathbf{G}_i$, and (3) α supports S.

Epistemic justification has to do with the *current* status of a proposition. As reasoning progresses, a sequent can fluctuate repeatedly between being justified and being unjustified. For the purpose of evaluating a system of defeasible reasoning, it is useful to be able to talk about its behavior "in the limit". The notion of a sequent's being justified in the limit is what was called "warrant" in chapter 2. The definition of "warrant" given there can now be stated more precisely:

> A sequent S is *warranted to degree* δ iff there is an i such that for every $j \geq i$, S is justified to degree δ at stage j.

Warrant, in this sense, characterizes the limit to which epistemic justification tends as reasoning proceeds.

Warrant is one way of understanding "justification in the limit". However, there is another obvious way of understanding this expression. Let $\mathbf{G}$ be the set of all nodes produced or producible by the reasoner, that is, $\bigcup_{i \in \omega} \mathbf{G}_i$. We can define a sequent to be *ideally warranted* iff it is justified relative to the whole set $\mathbf{G}$. More precisely:

> A sequent S is *ideally warranted to degree* δ *(relative to* ***input****)* iff there is a node α (relative to ***input***) of strength greater than or equal to δ such that (1) $\alpha \in \mathbf{G}$, (2) α is undefeated relative to $\mathbf{G}$, and (3) α supports S.

Ideal warrant has to do with what a reasoner should believe if it could produce all possible relevant arguments and then survey them. What is the relationship between warrant and ideal warrant? Interestingly, they need not be the same:

(10.1) Warrant does not entail ideal warrant.

Proof: Suppose there is an infinite sequence $\alpha_1,\ldots,\alpha_i,\ldots$ of nodes such that for each i, α_{i+1} defeats α_i and for each odd i, α_{i+1} is produced by the reasoner before α_i. Then α_1 is undefeated relative to every $\mathbf{G}_i$ containing α_1, so its conclusion is warranted. But relative to $\mathbf{G}$, α_1 is provisionally defeated, and hence its conclusion is not ideally warranted. To illustrate the possibility of such an infinite sequence of nodes, we can postulate an infinite sequence of propositions Q_i all either in ***input*** or supported by reasoning, and such that if we define D_i recursively by stipulating that $D_1 = (Q_0 \ldots P)$ and for $i > 1$, $D_{i+1} = (Q_i \ldots D_i)$, then (1) Q_0 is a prima

facie reason for P, and (2) for each $i \geq 1$, Q_i is a prima facie reason for D_i.

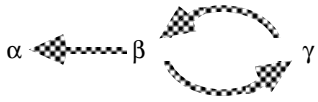
(10.2) Ideal warrant does not entail warrant.
Proof: Suppose there is a node α and an infinite set of pairs of nodes $\langle \beta_i, \gamma_i \rangle$ such that for each i, β_i defeats α and γ_i defeats β_i. Suppose further that each β_i and γ_i is undefeated at the stage of reasoning at which it is first produced. Then α will cycle between being defeated outright relative to G_i and undefeated relative to G_i for progressively larger i, but α is defeated relative to G. Thus the conclusion of α is ideally warranted but not warranted.

Warrant represents the "actual" behavior of the reasoner in the limit. We can think of ideal warrant as representing the set of conclusions we would "like" the reasoner to draw in the limit. Ideal warrant is a kind of ideal target. One kind of ideal adequacy condition for a defeasible reasoner is that these concepts should pick out the same conclusions. Let us define:

> The sequence $\{\alpha_i\}_{i \in \omega}$ of nodes is *ideal* iff for every sequent S, S is warranted iff it is ideally warranted.
>
> A defeasible reasoner is *ideal* iff the sequence of nodes it produces at progressively later stages of reasoning is ideal.

The preceding two theorems show that there can be argument sequences that are not ideal. Under what circumstances will a defeasible reasoner be ideal? In investigating this question, it is useful to think about defeat graphically, in terms of the defeat links that are represented in the inference graph. I will construe the directionality so that it is the parents of a node that defeat it. In diagraming defeat, I will write "α  β" when β is a parent of (defeats) α.

A *defeat branch* is any finite or infinite sequence $\{\xi_k\}$ such that for each i, $\langle \xi_i, \xi_{i+1} \rangle$ is a link. A *circular branch* is an infinite defeat-branch that repeats, i.e., there is a k such that for every $i \geq k$, there is a $j < k$ such that $\xi_i = \xi_j$. Collective defeat gives rise to circular branches. For example, the following collective defeat



gives rise to the circular branch $\langle\alpha,\beta,\gamma,\beta,\gamma,\beta,\gamma,\ldots\rangle$. Unless they are defeated by other nodes not on the branch, the nodes of a circular branch are all provisionally defeated relative to G.

For $\{\alpha_i\}_{i\in\omega}$ to be ideal, it is sufficient that for each node β, the status of β eventually stabilizes. Define:

> The argument sequence $\{\alpha_i\}_{i\in\omega}$ is *stable in the limit* iff for every argument β in G, β is undefeated (or defeated) relative to G iff there is a stage n such that for every $m \geq n$, β is undefeated (or defeated, respectively) relative to G_m.

Clearly, if $\{\alpha_i\}_{i\in\omega}$ is stable in the limit then it is ideal. The only way stability in the limit can fail is for the status of some node β to cycle indefinitely at progressively later stages of reasoning. This can happen only if there is an infinite subsequence $\{\xi_k\}_{k\in\omega}$ of $\{\alpha_i\}_{i\in\omega}$ all of whose members are relevant to the defeat-status of β. Obviously there are only two ways to get infinitely many nodes connected to β: (1) either β or some node that is a defeat ancestor of β could be an infinite defeat-branch-point in G; or (2) G could contain an infinite defeat-branch having β as its initial node. Let us consider these two possibilities separately.

Infinite defeat-branch-points (nodes defeated by infinitely many nodes) can lead to infinite cycling. This is illustrated by an example having the structure used in the proof of theorem (10.2). Suppose *input* = {*P*,*R*,*S*}, where *P* is a prima facie reason for *Q*, *R* is a prima facie reason for each of an infinite list of propositions D_i, each D_i is a prima facie reason for $(P \ldots Q)$, and *S* is a prima facie reason for each proposition $(D_i \ldots(P \ldots Q))$. With this set of prima facie reasons, if the ordering of the nodes is such that those supporting D_i and inferring $(P \ldots Q)$ from it alternate with those supporting $(D_i \ldots(P \ldots Q))$, then *Q* will alternate indefinitely between being justified and being unjustified at the different stages.

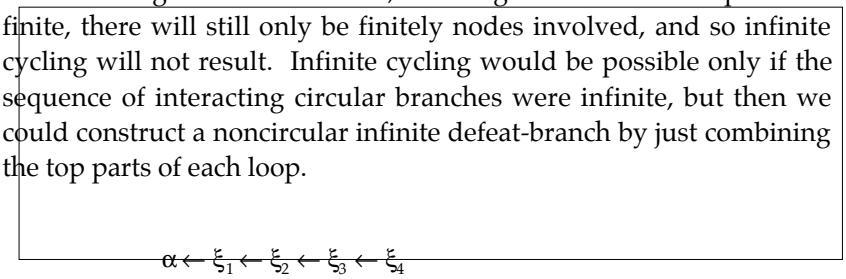
That the mere existence of infinite defeat-branch-points is insufficient to guarantee infinite cycling is obvious when we realize that, for many argument systems, it will be possible to construct infinitely many variants for any given argument. For example, we may be able to construct notational variants or add unnecessary steps. This can have the consequence that every defeat-branch-point will be an infinite defeat-branch-

point. But this need not give rise to infinite cycling because the different defeat-branches are not independent; anything defeating a node on one will defeat the corresponding node on the other. Let us define:

> A node β is *subsumed by* a node γ iff any defeater for γ or one of its inference ancestors is also a defeater for β or one of its inference ancestors.

A necessary (but not sufficient) condition for infinite cycling to result from an infinite defeat-branch-point is that there is no finite set of parents of the branch point such that every other parent is subsumed by one of the parents in the finite set. I will call such a branch point *a non-redundantly infinite defeat-branch-point*. If a defeat-branch-point is not non-redundantly infinite, then defeating a finite set of its parents would defeat them all, and so only by having infinite cycling at one of the parents could we get infinite cycling at the branch point.

Noncircular infinite defeat-branches can also lead to infinite cycling. For example, as in theorem (10.1), infinite cycling would result if there were an infinite sequence of propositions Q_i all either in *input* or supported by reasoning, and such that if we define D_i recursively by stipulating that $D_1 = (Q_0 \ldots P)$ and for $i > 1$, $D_{i+1} = (Q_i \ldots D_i)$, then (1) Q_0 is a prima facie reason for P, and (2) for each $i \geq 1$, Q_i is a prima facie reason for D_i. On the other hand, circular branches cannot lead to infinite cycling. They contain only finitely many different nodes, so once those nodes are all generated by the reasoner, they will be marked as provisionally defeated and will stay that way unless one of the nodes is defeated by other nodes not on the branch. If one of the nodes of the circular branch is defeated by a node ν not on the branch, that will lead to infinite cycling only if ν cycles infinitely. ν itself could be on another circular branch, and so on. Thus we might get a sequence of interacting circular branches, as in figure 15. If the sequence is finite, there will still only be finitely nodes involved, and so infinite cycling will not result. Infinite cycling would be possible only if the sequence of interacting circular branches were infinite, but then we could construct a noncircular infinite defeat-branch by just combining the top parts of each loop.



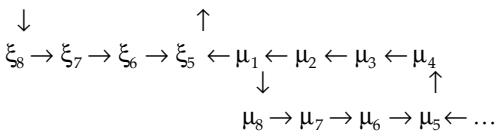


Figure 15. Interacting circular branches

Summarizing, we have the following simple lemma:

Lemma: If G contains no noncircular infinite defeat-branches and no non-redundantly infinite defeat-branch-points, then $\{\alpha_i\}_{i\in\omega}$ is stable in the limit, and hence ideal.

Although arrays of prima facie reasons and defeaters that will generate noncircular infinite defeat-branches and non-redundantly infinite defeat-branch-points are a formal possibility, I doubt that they are a real possibility. That is, we cannot find real examples of prima facie reasons having these structures. (My only reason for saying this is that I have tried and failed.) Accordingly, my strategy will be to adopt some realistic assumptions about the structure of the set of prima facie reasons that preclude these possibilities, and impose them as constraints on the design of a defeasible reasoner. It will then follow that the sequence of nodes produced by the reasoner is ideal.

Unless they are defeated by side branches, noncircular infinite defeat-branches lead to the provisional defeat of all of their nodes, but this defeat is peculiar because it does not arise from collective defeat. The only cases of provisional defeat that seem to arise in realistic systems of prima facie reasons and defeaters involve collective defeat. Accordingly, my first assumption will be that noncircular infinite defeat-branches are impossible:

Assumption 1: G contains no noncircular infinite defeat-branches.

The second assumption to be adopted is:

Assumption 2: For every proposition P and finite set X of propositions, there is a finite (possibly empty) set $nodes$ of nodes in G supporting P relative to X such that any other node in G that supports P relative to X is subsumed by some member of $nodes$.

Both of these assumptions are finiteness assumptions. They tell us that for any finite *input* and finite supposition X, there is a limit to how much non-deductive reasoning we can do. Again, the only reason for making these assumptions is that I can think of no plausible counterexamples. Both assumptions ought to be provable for particular classes of arguments. Note, however, that their truth will depend essentially on what arrays of prima facie reasons and defeaters are supplied for the use of the defeasible reasoner.

Assumption 1 precludes infinite cycling resulting from infinite defeat-branches. The role of assumption 2 is to rule out non-redundantly infinite defeat-branch-points. To see that it does this, consider any node α in G. α has only finitely many inference ancestors, and a node β can render α defeated only by defeating one of those ancestors. If a defeasible step infers P from Γ relative to a supposition X, then a node defeating this step must support either $\sim P$ or $(\Pi\Gamma \ldots P)$ relative to a subset of X. Accordingly, there are only finitely many conclusions that a defeating node can have, and hence by assumption 2, there is a finite set $nodes$ of nodes in G supporting those conclusions and such that any other node in G that supports one of those conclusions is subsumed by some member of $nodes$. In other words, non-redundantly infinite defeat-branch-points are impossible.

We now have the following theorem:

(10.3) A defeasible reasoner is ideal if assumptions 1 and 2 hold.

This follows from the fact that if assumptions 1 and 2 hold then G contains no non-circular infinite defeat-branches and no non-redundantly infinite defeat-branch-points, and hence by the lemma, $\{\alpha_i\}_{i\in\omega}$ is ideal. Theorem (10.3) will be the fundamental theorem making possible the construction of the defeasible reasoner in chapter four.

11. Logical Properties of Ideal Warrant

It follows from the analysis of ideal warrant that it exhibits a number of simple formal properties. Let us introduce the following symbolization:

$\mid\!\overset{\delta}{\Rightarrow}_{input} P$ iff P is ideally warranted to degree δ (relative to *input*).

For some purposes, it is useful to be able to talk about the *defeasible consequences* of a set of assumptions. These are the propositions that

can be inferred from those assumptions (relative to a fixed set *input*). Precisely:

> $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} P$ (*P is a defeasible consequence of* Γ *relative to* ***input***, and Γ *defeasibly implies P*) iff there is a node α of strength greater than or equal to δ such that (1) $\alpha \in \mathcal{G}$, (2) α is undefeated relative to $\mathcal{G}$, and (3) α supports *P* relative to Γ.

Ideal warrant and defeasible implication have a number of obvious logical properties. Define:

> $\Gamma \vdash P$ (*P is a deductive consequence of* Γ, and Γ *deductively implies P*) iff there is a node α such that (1) $\alpha \in \mathcal{G}$, (2) neither α nor any of its inference ancestors is a pf-node, (3) no member of ***input*** is an inference ancestor of α, and (4) α supports *P* relative to Γ.
>
> *P* and *Q* are *deductively equivalent* iff $\{P\} \vdash Q$ and $\{Q\} \vdash P$.

Then we have:

> $\mid\!\overset{\delta}{\Rightarrow}_{input} P$ iff $\varnothing \mid\!\overset{\delta}{\Rightarrow}_{input} P$.
>
> If $\Gamma \vdash P$ then $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} P$.
>
> If $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} P_1$ and ... and $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} P_n$, and $\{P_1,...,P_n\} \vdash Q$, then $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} Q$.
>
> If $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} P$ and $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} Q$, then $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} (P \& Q)$.
>
> If *P* and *Q* are deductively equivalent then $\Gamma \cup \{P\} \mid\!\overset{\delta}{\Rightarrow}_{input} R$ iff $\Gamma \cup \{Q\} \mid\!\overset{\delta}{\Rightarrow}_{input} R$.

If $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} P$ and $\Gamma \cup \{P\} \mid\!\overset{\delta}{\Rightarrow}_{input} Q$, then $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} Q$.

If $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} P$ and $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} Q$ then $\Gamma \cup \{P\} \mid\!\overset{\delta}{\Rightarrow}_{input} Q$.[12]

If $\Gamma \cup \{P\} \mid\!\overset{\delta}{\Rightarrow}_{input} Q$, then $\Gamma \mid\!\overset{\delta}{\Rightarrow}_{input} (P \supset Q)$.[13]

12. Conclusions

This completes the semantical theory of defeasible epistemic reasoning. This chapter has developed the accounts of justification and warrant based upon the argument-based approach to defeasible reasoning in terms of prima facie reasons, rebutting defeaters, and undercutting defeaters. The next task will be to construct a procedural theory of defeasible reasoning based upon this semantical theory. The result will be the automated defeasible reasoner described in chapter 4.

---

[12] This is the principle Gabbay [1985] calls "restricted monotonicity".

[13] For further discussion of such logical principles, see [Gabbay 1985, 1991], Kraus, Lehmann, and Magidor [1990], and Makinson [1988].