Self-Defeating Arguments

JOHN L. POLLOCK
Department of Philosophy, University of Arizona, Tucson, AZ 85721, U.S.A.
(e-mail: pollock @ ccit.arizona.edu)

Abstract. An argument is *self-defeating* when it contains defeaters for some of its own defeasible lines. It is shown that the obvious rules for defeat among arguments do not handle self-defeating arguments correctly. It turns out that they constitute a pervasive phenomenon that threatens to cripple defeasible reasoning, leading to almost all defeasible reasoning being defeated by unexpected interactions with self-defeating arguments. This leads to some important changes in the general theory of defeasible reasoning.

Key words. Argument, defeasible, nonmonotonic.

1. Introduction

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. The articulation of the logical structure of defeasible reasoning has become an important topic in both philosophical epistemology and artificial intelligence. My ultimate objective is twofold – to construct a philosophical theory of defeasible reasoning, and to build an automated reasoner that implements the philosophical theory in a computer system. This task proves to be unexpectedly difficult, partly because most work on the structure of reasoning has focused exclusively on deductive reasoning, and there are some surprising phenomena that occur only in connection with defeasible reasoning. This paper is concerned with one such phenomenon. I first noted the existence of self-defeating arguments several years ago (Pollock, 1987), but at this time I took them to be an isolated phenomenon of little more than technical interest. I have subsequently come to realize that they represent a pervasive phenomenon that would cripple defeasible reasoning without some adequate mechanism for handling them. That is the topic of this paper. The phenomenon of self-defeating arguments arises against the background of the theory of defeasible reasoning that I have developed elsewhere, so I will begin by sketching that theory. The theory is of the general sort that Lin and Shoham (1989) call *the argument-based approach*. The early development of my theory is recapitulated in Pollock (1974 and 1986), and the more recent developments are in Pollock (1987, 1990, 1990a, and 1991). Having sketched the theory of defeasible reasoning, I will display the phenomenon that constitutes my subject matter, and investigate the changes it forces in my theory. The present paper builds upon several earlier papers (particularly Pollock, 1987 and 1991), but is intended to be essentially self-contained.

Minds and Machines **1**: 367–392, 1991.

2. Prima Facie Reasons and Defeaters

Reasoning proceeds by constructing arguments, where *reasons* provide the atomic links in arguments. *Conclusive reasons* logically entail their conclusions. Defeasibility arises from the fact that not all reasons are conclusive. Those that are not are *prima facie reasons*. Prima facie reasons create a presumption in favor of their conclusion, but it is defeasible. For example, something's looking red to me provides a prima facie reason for thinking that it is red. If I have no other relevant information, this makes it reasonable for me to believe that the object is red, but if I also have some independent good reason for thinking that the object is not red, that defeats the prima facie reason.

I will take reason to be an ordered pair $\langle \Gamma, p \rangle$ where Γ is the set of premises of the reason and p is the conclusion. The simplest kind of defeater for a prima facie reason $\langle \Gamma, p \rangle$ is a reason and p is 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. Experience in using prima facie reasons in epistemology indicates that there are no such prima facie reasons.[1] Every prima facie reason has associated defeaters that are not rebutting defeaters, and these are the most important kinds of defeaters for understanding any complicated reasoning. Defeaters that are not rebutting defeaters attack a prima facie reason without attacking its conclusion. They accomplish this by instead attacking the connection between the premises and the conclusion. For instance, $\ulcorner x$ looks red$\urcorner$ is a prima facie reason for $\ulcorner x$ is red$\urcorner$. 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, $\ulcorner x$ is illuminated by red lights and red lights can make things look red when they are not$\urcorner$ 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 $\ulcorner x$ looks red$\urcorner$ and $\ulcorner x$ is red$\urcorner$, giving us a reason for doubting that x wouldn't look red unless it were red. $\ulcorner P$ wouldn't be true unless Q were true $\urcorner$ is some kind of conditional, and I will symbolize is as $\ulcorner P \gg Q \urcorner$. The preceding indicates that 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 $\ulcorner \Pi\Gamma \gg p \urcorner$ is a defeater. I call these *undercutting defeaters*:

> 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 $q = \ulcorner \sim (\Pi\Gamma \gg p) \urcorner$.

A prima facie reason to which I will appeal repeatedly can be formulated roughly as follows:

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

The simplest undercutting defeaters for the statistical syllogism are *subproperty defeaters*:

$\ulcorner Hc \ \& \ \text{prob}(F/G\&H) < \text{prob}(F/G) \urcorner$

is an undercutting defeater for the above.

The statistical syllogism and its defeaters are discussed at great length in Pollock (1990), and the reader is referred there for more details.

3. Arguments

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. Reasons are combined in various patterns to form arguments. The simplest arguments are *linear* arguments. These can be viewed as finite sequences of propositions each of which is either a member of *input* or inferable from previous members of the sequence in accordance with some reason schema.

It is important to realize that not all arguments are linear. The easiest way to see this is to note that linear arguments 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).

In suppositional reasoning, we can no longer think of arguments as finite sequences of propositions, because each line of an argument may depend upon

suppositions. We can instead think of lines of arguments as ordered triples $\langle X, p, \beta \rangle$ where X is the set of propositions comprising what is supposed on that line, p is the proposition obtained on that line, and β describes the *basis* for the line. β will be taken to be an ordered pair $\langle \lambda, R \rangle$ where R is the rule of inference used to obtain the present line and λ is the set of line numbers of the lines from which the present line is inferred by using R. X is the *supposition set* of the line. Linear arguments can be viewed as arguments in which the supposition sets are always empty. Discharge rules are rules that manipulate supposition sets. For instance, *conditionalization* could be formulated as follows:

$$\text{From } \langle X \cup \{p\}, q, \beta \rangle, \text{ infer } \langle X, (p \supset q), \langle \{i\}, \textit{conditionalization} \rangle \rangle .$$

Rules of inference are really rules for the construction of arguments, so *conditionalization* can be stated more precisely by explicitly construing it as such. Given a finite sequence σ, let σ_i be the ith element of σ, and let $\sigma^\frown x$ be the result of appending x to the end of σ. *Conditionalization* can then be stated as follows:

> If σ is an argument and $\sigma_i = \langle X \cup \{p\}, q, \beta \rangle$, then $\sigma^\frown \langle X, (p \supset q), \langle \{i\}, \textit{conditionalization} \rangle \rangle$ is also an argument.

Other rules for argument formation will include the following:

Input
If $p \in \textit{input}$ and σ is an argument, then for any X, $\sigma^\frown \langle X, p, \langle \varnothing, \textit{input} \rangle \rangle$ is an argument.

Supposition
If σ is an argument and X is any finite set of propositions, then if $p \in X$, $\sigma^\frown \langle X, \{p\}, p, \langle \varnothing, \textit{supposition} \rangle \rangle$ is an argument.

Reason
If σ is an argument, $\langle X, p_1, \beta_1 \rangle, \ldots, \langle X, p_n, \beta_n \rangle$ are the $i_1 - i_n$ lines of σ, $\langle \{p_1, \ldots, p_n\}, q \rangle$ is a reason (either conclusive or prima facie), then $\sigma^\frown \langle X, q, \langle \{i_1, \ldots, i_n\}, \textit{reason} \rangle \rangle$ is an argument.

Dilemma
If σ is an argument and $\sigma_i = \langle X, (p \vee q), \beta_i \rangle$, $\sigma_j = \langle X \cup \{p\}, r, \beta_j \rangle$, and $\sigma_k = \langle X \cup \{q\}, r, \beta_k \rangle$ then $\sigma^\frown \langle X, r, \{i, j, k\}, \textit{dilemma} \rangle \rangle$ is an argument.

Other rules of argument formation should be included as well, but these will suffice for the present discussion. An argument σ *supports* the proposition p *relative to* the supposition X iff for some i and β, $\sigma_i = \langle X, p, \beta \rangle$. σ *supports* p iff σ supports p relative to the empty supposition. The *conclusion* of an argument is its last line.

4. Reasoning and Warrant

A proposition is *warranted* in a particular epistemic situation iff, starting from that epistemic situation, an ideal reasoner unconstrained by time or resource limitations will eventually reach a point where it believes the proposition and will never subsequently retract it. Warranted propositions are those that would be justified "in the long run" if the reasoner were able to do all possible relevant reasoning. A characterization of the set of warranted propositions can be given fairly easily if we take as primitive the notion of one argument defeating another. Suppose we have an argument α supporting a conclusion P, and an argument β that defeats α. If these are the only relevant arguments, then P is not warranted. But now suppose we acquire a third argument γ that defeats β. This situation is diagrammed in Figure 1. The addition of γ should have the effect of reinstating α, thus making P warranted. We can capture this kind of interplay between arguments by talking about arguments being *in* or *out* at different *levels*. Let us (provisionally) define:

> All arguments are *in at level* 0.
>
> An argument is *in at level* $n+1$ iff it is in at level 0 and it is not defeated by any argument in at level n.

Thus α, β, and γ are all in at level 0. γ is in at level 1, but neither α nor β is in at level 1. Accordingly, α and γ are in at level 2, but β is not. And for every $n \geqslant 2$, α and γ are in at level n, but β is out. Let us define:

> An argument is *ultimately undefeated* iff there is an m such that for every $n \geqslant m$, the argument is in at level n.

My initial proposal is then that a proposition is warranted iff it is supported by some ultimately undefeated argument.[2] It will be suggested below that this proposal must be modified slightly, but this characterization will suffice for now.

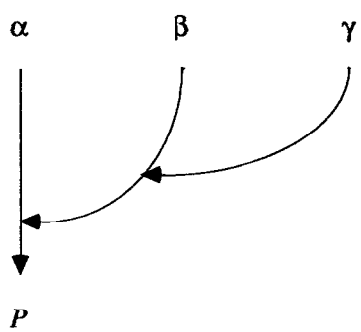
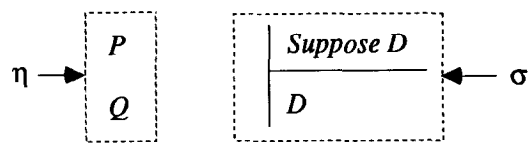


Fig. 1. Interacting arguments.

5. Defeat among Arguments

I have characterized warrant in terms of arguments being in or out at different levels, where the latter notion is defined in terms of one argument defeating another. To complete the theory of warrant, we must characterize when arguments defeat one another. A general treatment of defeat among arguments involves addressing a complex issue that has rarely been addressed in either philosophy or AI. Reasons differ in strength. Some reasons are better than others. If we have a reason for p and a reason for $\neg p$, but the latter is significantly stronger than the former, then it wins the competition and we should believe $\neg p$. Thus a general theory of reasoning requires us to talk about the strengths of reasons and how those strengths affect interactions between reasons. However, I am *not* going to address that problem here. I have addressed it in another paper (Pollock, 1991). Unfortunately, I now feel that the account given in that paper was inadequate for reasons unrelated to the strengths of arguments. The purpose of this section of the present paper is to correct those inadequacies. Those inadequacies arise even if we adopt the simplifying assumption that all reasons are of the same strength. Accordingly, that will be the assumption that is made here. The account given here can be smoothly integrated into the earlier theory of reasons of varying strengths, but I will not explicitly discuss that here.

An argument σ defeats an argument η by supporting a defeater for some line of η. Recall, however, that in suppositional reasoning, different lines of an argument may depend upon different suppositions. If a prima facie reason $\langle \Gamma, p \rangle$ is used in η, the presence in σ of a defeater for this reason does not automatically guarantee that σ defeats η. 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:



The defeater D is introduced into σ as a mere supposition, and that supposition is not included in the supposition set of the line of η on which the prima facie reason $\langle \{P\}, Q \rangle$ is used. Clearly, we should not be able to defeat an argument just by *supposing* a defeater that has no independent justification. It seems that the occurrence in σ of a defeater on a line whose supposition set is X should only defeat a use of $\langle \Gamma, p \rangle$ in η on a line whose supposition set includes X. Accordingly, we can define:

An argument σ *rebuts* an argument η iff:
(1) some line of η has the form $\langle Y, q, \langle \zeta, \textit{reason} \rangle \rangle$ where the propositions supported on the lines in ζ constitute a prima facie reason for q; and
(2) the last line of σ has the form $\langle X, \neg q, \beta \rangle$ where $X \subseteq Y$.

An argument σ *undercuts* an argument η iff:

(1) some line of η has the form $\langle Y, q, \langle \zeta, \textit{reason} \rangle \rangle$ where the propositions $p_1, \ldots, p_k$ supported on the lines in ζ constitute a prima facie reason for q; and

(2) the last line of σ has the form $\langle X, \sim((p_1 \,\&\, \ldots \,\&\, p_k) \Rightarrow q), \beta \rangle$ where $X \subseteq Y$.

An argument σ *directly defeats* an argument η iff σ either rebuts or undercuts η.

Then it seems reasonable to propose:

An argument σ *defeats* an argument η iff a subargument of σ directly defeats η.

It turns out, however, that an additional source of defeat must be recognized, and a refinement made to the analysis of warrant. These will be the topics of the next several sections. To explain the need for these refinements to the theory, I must first discuss the phenomenon of collective defeat.

6. Collective Defeat

In simple cases, an argument σ is defeated by there being another argument that warrants a defeater for a defeasible step of σ. But in more complicated cases, arguments can be defeated without any of their defeaters being warranted. This is the phenomenon of *collective defeat* wherein arguments are defeated collectively rather than individually. Consider a simple scenario in which *input* $= \{p, q\}$, and $\langle \{p\}, r \rangle$ and $\langle \{q\}, \neg r \rangle$ are prima facie reasons of the same strength. Then we can construct two simple arguments as in Figure 2. It follows from our analysis that each argument defeats the other. Accordingly, they are in at level 0, out at level 1, in again at level 2, out again at level 3, and so on. Hence neither is ultimately undefeated, and hence neither r nor $\neg r$ is warranted.

Collective defeat operates in accordance with the following general principle:

Fig. 2. Collective defeat.

THE PRINCIPLE OF COLLECTIVE DEFEAT. If Σ is a set of arguments such that (1) each argument in Σ is defeated by some other argument in Σ, and (2) no argument in Σ is defeated by any argument not in Σ, then no argument in Σ is ultimately undefeated.

This is because each argument in Σ will be in at every even level, but then it follows that each will be out at every odd level. We can define:

> An argument σ is *defeated outright* iff there is a level n such that σ is out at all higher levels.
>
> An argument σ is *provisionally defeated* iff there is no level n such that σ is in at all higher levels or out at all higher levels.

A proposition *undergoes provisional defeat* iff some arguments supporting it are provisionally defeated and any other arguments supporting it are defeated outright.

Collective defeat can be fruitfully illustrated by a problem that has plagued many theories of probabilistic reasoning. This is the lottery paradox.[3] Suppose you hold one ticket in a fair lottery consisting of one million tickets, and suppose it is known that one and only one ticket will win. Observing that the probability is only 0.000001 of a ticket 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, namely, 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 warranted in believing of any ticket that it will not win.[4] This is captured formally by the principle of collective defeat, which tells us that our prima facie reasons collectively defeat one another.

Collectively defeated arguments are provisionally defeated, but it turns out that an argument can be provisionally defeated without entering into collective defeat with other arguments. This turns upon the fact that although an argument that is defeated outright cannot defeat another argument, provisionally defeated arguments can still provisionally defeat other arguments. To illustrate, suppose α and β defeat one another collectively. In this case, α is in at every even level, and out at every odd level. Now suppose α supports a defeater for a third argument γ. This will have the effect that γ is out at every odd level, and back in at every even level, so γ is also provisionally defeated, even though it may not defeat the arguments at whose hands it suffers provisional defeat. Consequently, γ may be provisionally defeated without entering into collective defeat with any other arguments.

7. Directly Self-Defeated Arguments

Armed with an understanding of collective defeat, the need for an additional source of defeat among arguments 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 $input = \{P, Q, S\}$. Then we can construct the following three arguments (where defeasible inferences are indicated by dashed arrows):

$$\alpha \quad P \dashrightarrow R \qquad \beta \quad Q \dashrightarrow \sim R \qquad \sigma \quad S \dashrightarrow T$$

α and β collectively defeat one another, but σ should be independent of α and β and ultimately undefeated. The difficulty is that we can construct a fourth argument (where deductive inferences are indicated by solid arrows):

$$\eta \quad \left.\begin{array}{l} P \dashrightarrow R \longrightarrow (R \vee \sim T) \longrightarrow \\ Q \dashrightarrow \sim R \longrightarrow \end{array}\right\} \longrightarrow \sim T$$

η uses a standard strategy for deriving an arbitrary conclusion from a contradiction. The problem is now that η rebuts σ. Of course, η itself is defeated by either α or β, or for that matter, by itself (it supports defeaters for its own defeasible steps). But that only results in η being collectively defeated, and as I pointed out above, a collectively defeated argument can still provisionally defeat another argument. η is out at every even level, but it is still in at every odd level. Consequently, it still forces σ to be out at every even level, and hence σ is provisionally defeated too. But σ should not be provisionally defeated – it should be ultimately undefeated.

To avoid this difficulty, η must be out at every level, not just at every even level. In other words, η must undergo more than just provisional defeat. There is no way to get this result from the analysis of defeat proposed above. My diagnosis of the difficulty is that η is "internally defective". η is *directly self-defeated* in the sense that it supports defeaters for some of its own defeasible steps. By the proposal of the previous section, this means that it enters into collective defeat with itself, but my suggestion is that this should be regarded as a more serious defect – one which removes it from competition with other arguments altogether (and hence leaves σ undefeated). Let us define:

η is *directly self-defeated* iff η defeats itself.

The phenomenon of direct self-defeat can be further illustrated by looking at what appears to be a paradox of defeasible reasoning. This concerns the lottery paradox again. The lottery paradox is generated by supposing that a proposition R describing the lottery (it is a fair lottery, has one million tickets, and so on) is warranted. Given that R is warranted, we get collective defeat for the proposition

that any given ticket will not be drawn. But the present account makes it problematic how R can be warranted. Normally, we will believe R on the basis of being told that it is true. In such a case, our evidence for R proceeds in accordance with the statistical syllogism. That is, we know inductively that most things we are told are true, and that gives us a prima facie reason for believing R. So we have only a defeasible reason for believing R. Let σ be the argument supporting R. Let T_i be the proposition that ticket i will be drawn. In accordance with the standard reasoning involved in the lottery paradox, we can extend σ to generate a longer argument η supporting $\sim R$. This is diagrammed in Figure 3. The final step of the argument proceeds by noting that the $\sim T_i$ jointly entail $\sim R$, because if none of the tickets is drawn then the lottery is not fair. Thus we generate the argument η of Figure 3. The difficulty is now the η rebuts σ. Thus by the proposal of Section 4.1, σ and η defeat one another, with the result that neither is ultimately undefeated. In other words, R undergoes collective defeat. Again, this result is intuitively wrong. It should be possible for us to become warranted in believing R on the basis described. I propose once more than the solution to this problem lies in noting that because η contains σ, η is directly self-defeated.

Both of the preceding difficulties can be avoided by ruling that directly self-defeated arguments are defeated absolutely – not just provisionally. As they are defeated absolutely, they cannot enter into collective defeat with other arguments, and so the arguments σ of the preceding two examples are ultimately undefeated, as they should be. This can be accomplished by ruling that directly self-defeated arguments are out at every level. The simplest way to accomplish this is to revise the conditions under which an argument is in at level 0:

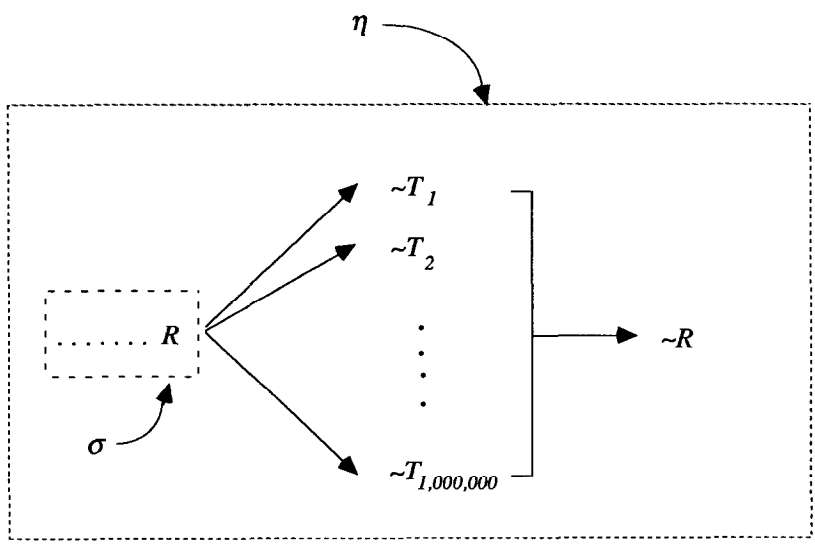


Fig. 3. The lottery paradox.

An argument η is *in at level* 0 iff η is not directly self-defeated.

In order to be in at level k, an argument must be in at level 0, so this ensures that η is out at every level.

8. Indirect Defeat

There is more to the story of self-defeated arguments than has emerged so far. The phenomenon of direct self-defeat forces us to recognize a new source of defeat among arguments. Directly self-defeated arguments are out at every level. It is, however, insufficient to just say this. Consider an argument η whose conclusion is unwarranted. Suppose the conclusion of η is obtained deductively from earlier steps. Then at least one of those earlier steps must also be unwarranted. Let us say that the *nearest defeasible ancestors* of an argument α are the subarguments of α whose last steps are defeasible and such that α is a deductive extension of those subarguments. If the last step of α is defeasible, then α is its own sole nearest defeasible ancestor. By the above reasoning:

THE PRINCIPLE OF SUBARGUMENT DEFEAT. If an argument is defeated, at least one of its nearest defeasible ancestors must be defeated.

In the two examples of directly self-defeated arguments considered in the previous section, this principle is automatically satisfied because the nearest defeasible ancestors collectively defeat one another. But there can be directly self-defeated arguments with more complicated structures in which this is no longer the case. To get clear on this, let us first distinguish between directly self-defeat*ing* arguments and directly self-defeat*ed* arguments:

> An argument α is *directly self-defeating* iff it is directly self-defeated but no proper subargument of α is directly self-defeated.

In other words, directly self-defeating arguments are those in which the self-defeat occurs first at the final step. Directly self-defeated arguments are then any extensions of directly self-defeating arguments. There are three different ways that an argument α can be directly self-defeating:

(1) α's conclusion is the result of a defeasible inference, and the conclusion of one of α's subarguments is a defeater for that inference;
(2) α's conclusion is inferred from the conclusions of two or more subarguments, and one of those subarguments defeats another;
(3) α's conclusion is a defeater for one of α's subarguments.

Case (1) is unproblematic, because if the earlier subargument is undefeated then α is defeated. Case (2) is similarly unproblematic. Suppose the conclusion of α is

inferred deductively from the conclusions of subarguments β and γ, and β supports a defeater for some line of γ. If β is undefeated then it follows that γ is defeated. Case (3), however, is more problematic. Suppose that the conclusion of α is a defeater for one of its own earlier lines, α_i. As α is defeated and it is inferred deductively from its nearest defeasible ancestors, the conclusions of some of those nearest defeasible ancestors must be defeated. However, nothing in our account to date ensures that. It is true that α defeats the subargument β whose conclusion is α_i, and hence defeats any of its nearest defeasible ancestors that contain β, however α is directly self-defeating and hence out at every level, so this does not result in defeat for β. We need a new principle ruling that β is defeated. To illustrate, suppose we know (1) that people generally tell the truth, (2) that Robert says that the elephant beside him looks pink, and (3) that Robert becomes unreliable in the presence of pink elephants. I assume that $\ulcorner x$ looks pink$\urcorner$ is a prima facie reason for $\ulcorner x$ is pink$\urcorner$. 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 4. For each n, let A_n be the argument whose conclusion is the nth proposition (numbered as in the figure). A_7 is directly self-defeating, so one of its

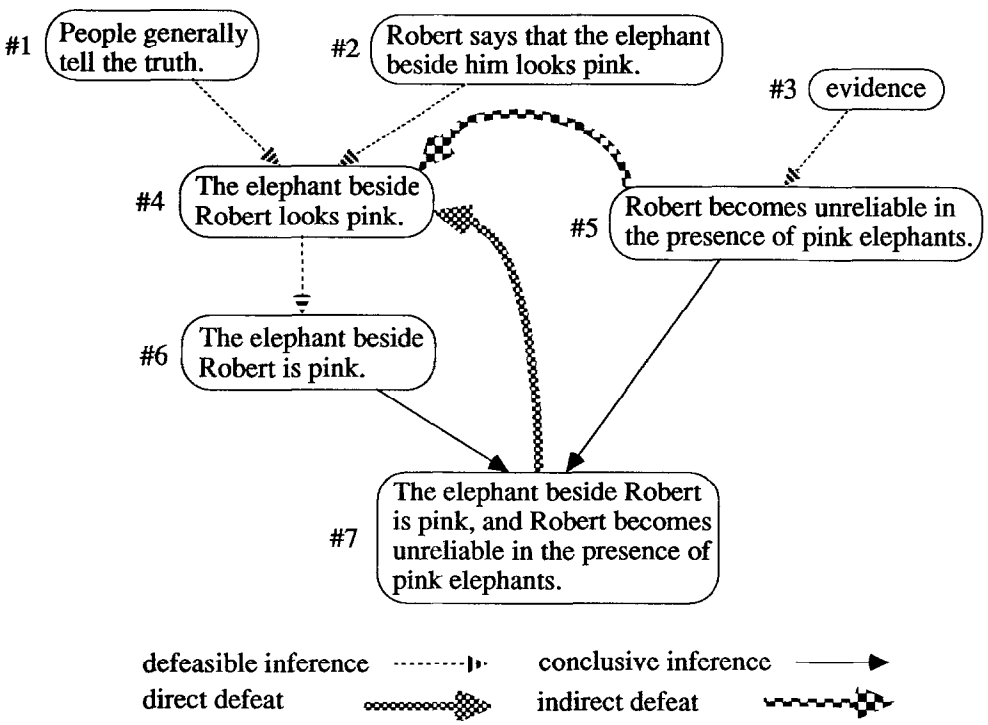


Fig. 4.

nearest defeasible ancestors must be defeated. These are A_5 and A_6. Of these, it seems clear that A_6 should be defeated *by* having A_4 defeated, because A_6 is the only nearest defeasible ancestor having A_4 as an ancestor, and it is the defeat of A_4 by A_7 that makes A_7 self-defeating. However, the fact that A_7 defeats A_4 is not enough to get the latter defeated, because A_7 is directly self-defeating and hence defeated itself. Thus there must be an independent source of defeat for A_4. Notice, however, that if A_5 becomes defeated in some way, then A_4 no longer needs to be defeated (because then A_7 will have a defeated subargument anyway). Thus A_4 must be defeated if A_5 is not, but not otherwise. This can be captured by saying that A_5 defeats A_4. On the other hand, A_5 does not support an undercutting or rebutting defeater for A_4, so if we are going to regard it as defeating A_4, this must illustrate a new kind of defeat. Let us call it *indirect defeat*. The general case in which indirect defeat arises is diagrammed in Figure 5. Here the conclusion R of the directly self-defeating argument defeats its own ancestor P. Let K be the set of all the nearest defeasible ancestors that do not contain the defeated subargument for P as one of their own subarguments. Then the members of K should jointly defeat P. Note that indirect defeat differs from rebutting and undercutting defeat in that it may involve several arguments *jointly* defeating another argument. It is not just a relationship between individual arguments.

Indirect defeat is made precise as follows:

> A set K of arguments *indirectly defeats* an argument β iff there is an argument α such that (1) α directly defeats β, and (2) K is the set of all the nearest defeasible ancestors of α that do not have β as a subargument.

The recognition of indirect defeat forces us to complicate some of our earlier

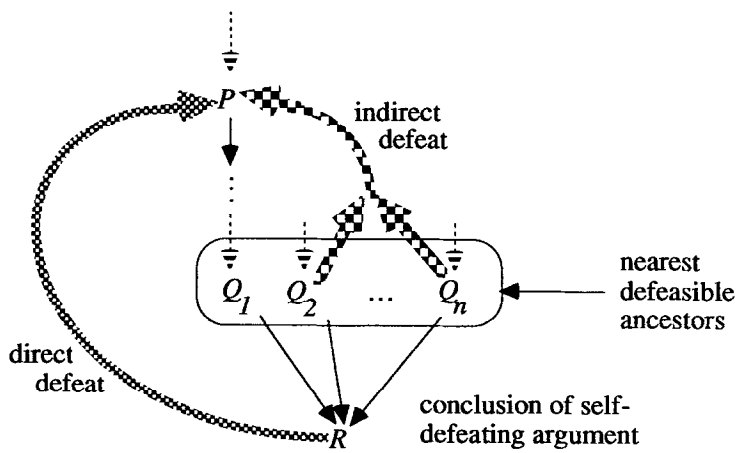


Fig. 5.

definitions. First, the definition of 'defeat among arguments' must be revised. The general strategy is to say that a set K of arguments *minimally defeats* an argument η iff the conclusion of the arguments in K do the defeating, and then define:

> An argument σ *defeats* an argument η iff a set of subarguments of σ minimally defeats η.

What remains is to define minimal defeat. We have uncovered two sufficient conditions for minimal defeat:

(1) If σ directly defeats η then $\{\sigma\}$ minimally defeats σ.

(2) If K indirectly defeats η then K minimally defeats η.

However, these sufficient conditions are not jointly necessary. Consider the scenario diagrammed in Figure 6. The arguments for Q_2 and Q_3 together defeat the argument for S by defeating its subargument for P. But notice that an argument supporting $\sim Q_2$ would also defeat the argument for S because it would defeat the subargument for Q_2. So it seems intuitively that Q_2 should be irrelevant to the defeat of the argument for S. The latter argument should be defeated by the argument for Q_3 all by itself. This should be regarded as another instance of minimal defeat. So we should have:

> If the union of K and a nonempty set of subarguments of η indirectly defeats a subargument of η, then K minimally defeats η.

Furthermore, the indirect defeat in this example could be replaced by any minimal defeat, indicating that minimal defeat satisfies the following recursive clause:

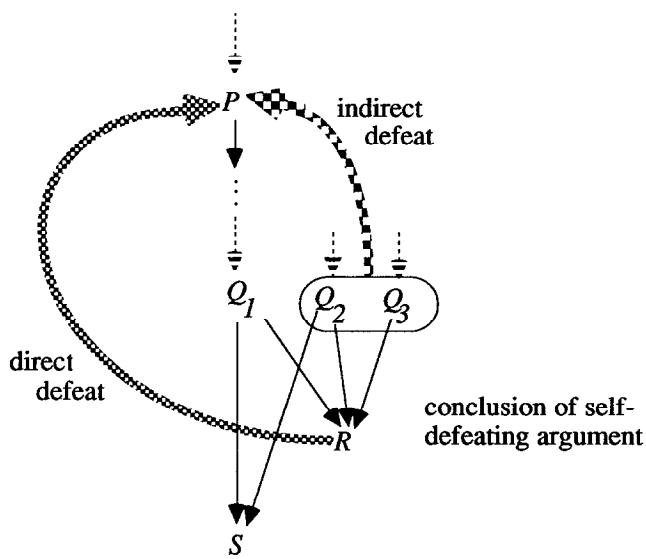


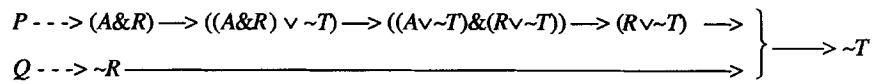
Fig. 6.

> If the union of K and a nonempty set of subarguments of η minimally defeats a subargument of η, then K minimally defeats η.

Accordingly, minimal defeat should be defined recursively as follows:

A nonempty set K of arguments *minimally defeats an argument* η iff either:
(1) for some σ, $K = \{\sigma\}$ and σ directly defeats η; or
(2) K indirectly defeats η; or
(3) the union of K and a nonempty set of subarguments of η minimally defeats a subargument of η.

The extended concept of defeat must be used in the definition of 'directly self-defeated argument'. This can be illustrated by modifying the initial example of section 7 by supposing that P is a prima facie reason for $(A\&R)$ rather than R alone. Then revise η as follows



This argument should be regarded as directly self-defeating, but it does not contain explicit defeaters for any of its lines. Instead, defeaters for either of the two prima facie inferences can be inferred deductively from the other lines. In particular, the argument α:

$$P \dashrightarrow (A\&R) \rightarrow R$$

defeats the subargument η_0:

$$Q \dashrightarrow \sim R$$

of η. It follows that the nearest defeasible ancestors of α indirectly defeat η_0. The only nearest defeasible ancestor is the argument for $(A\&R)$, so it follows that η is directly self-defeating.

The phenomenon of indirect defeat requires a further generalization of the concept of self-defeat. A special case of indirect defeat occurs when K is the empty set. This occurs when all of the nearest defeasible ancestors of α have β as a subargument. In this case, those nearest defeasible ancestors should be out at every level. This can be captured by saying that they are *indirectly self-defeating*:

> An argument η is *indirectly self-defeating* iff η is a nearest defeasible ancestor of an argument α such that (1) α directly defeats its own subargument β, and (2) β is a subargument of all of α's nearest defeasible ancestors.

This is illustrated by the argument η of Figure 3 (the lottery paradox). Let us combine these two concepts of self-defeat into a single concept:

An argument is *self-defeating* iff it is either directly self-defeating or indirectly self-defeating.

An argument is *self-defeated* iff it has a self-defeating subargument

The phenomenon of self defeat and indirect defeat turns out to be extremely important in understanding defeasible reasoning. The next two sections of the paper provide illustrations of this.

9. The Paradox of the Preface

I first noticed indirect defeat in connection with *the paradox of the preface*. In Pollock (1990), I presented the paradox of the preface 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."
>
> This is the paradox of the preface (so named because in the original version the author confesses in the preface that this book probably contains a falsehood).[5]

The paradox of the preface 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 are not warranted, whereas in the paradox of the preface we want to insist that the members of Γ are warranted. How can we explain the difference?

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 seems like an esoteric paradox of little more than theoretical interest. However, I have recently come to realize that 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 it surely 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 warranted. And as every color judgment is a member of some such set of 10,000, it follows that all color judgments are unwarranted. 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.

My proposal is that the paradox of the preface can be resolved by appealing to indirect defeat.[6] The paradox has the following form. We begin with a set $\Gamma = \{p_1, \ldots, p_N\}$ of propositions, where Γ has some property B (being the propositions asserted in a book of a certain sort, or being a set of propositions supported by arguments employing a certain prima facie reason), and we know that it is highly probable that such a set of propositions contains at least one false member. Letting T be the property of being true, we can express this probability as:

$$\text{prob}((\exists z)(z \in X \;\&\; \sim Tz)/B(X)) = r\,.$$

This high probability, combined with the premise $B(\Gamma)$, gives us 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 diagrammed in Figure 7 for each $\sim Tp_i$. Call this argument η_i.

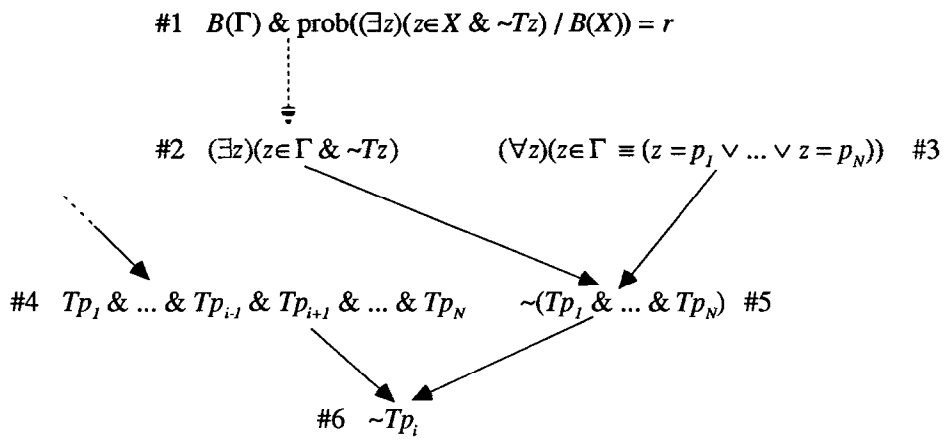


Fig. 7.

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

$$C(\Gamma) \ \&\ \text{prob}((\exists z)(z \in X \ \&\ \sim Tz)/B(X) \ \&\ C(X)) < r\,.^{7}$$

It is shown in Pollock (1990, p. 251) that

$$\begin{aligned}
&\text{prob}((\exists z)(z \in X \ \&\ -Tz)/B(X) \ \&\ X \\
&\quad = \{x_1, \ldots, x_N\} \ \&\ x_1, \ldots, x_N \text{ are distinct } \&\ (\forall z)(z \in X \\
&\quad \equiv (z = x_1 \vee \ldots \vee z = x_N)) \ \&\ Tx_1 \ \&\ \ldots \ \&\ Tx_{i-1} \ \&\ \ldots \ \&\ Tx_N) \\
&= \text{prob}(\sim Tx_i/B(X) \ \&\ X \\
&\quad = \{x_1, \ldots, x_N\} \ \&\ x_1, \ldots, x_N \text{ are distinct } \&\ (\forall z)(z \in X \\
&\quad \equiv (z = x_1 \vee \ldots \vee z = x_N)) \ \&\ Tx_1 \ \&\ \ldots \ \&\ Tx_{i-1} \ \&\ \ldots \ \&\ Tx_N)\,.
\end{aligned}$$

It is at this point that the paradox of the preface differs 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,

$$\begin{aligned}
&\text{prob}(\sim Tx_i/B(X) \ \&\ X \\
&\quad = \{x_1, \ldots, x_N\} \ \&\ x_1, \ldots, x_N \text{ are distinct } \&\ (\forall z)(z \in X \equiv (z = x_1 \\
&\quad \vee \ldots \vee z = x_N)) \ \&\ Tx_1 \ \&\ \ldots \ \&\ Tx_{i-1} \ \&\ Tx_{i+1} \ \&\ \ldots \ \&\ Tx_N) \\
&\leq \text{prob}(\sim Tx_i/B(X) \ \&\ X \\
&\quad = \{x_1, \ldots, x_N\} \ \&\ x_1, \ldots, x_N \text{ are distinct } \&\ (\forall z)(z \in X \\
&\quad \equiv (z = x_1 \vee \ldots \vee z = x_N)))\,.
\end{aligned}$$

There is no reason to believe that the condition $\ulcorner X = \{x_1, \ldots, x_N\}$ & $x_1, \ldots, x_N$ are distinct & $(\forall z)(z \in X \equiv (z = x_1 \vee \ldots \vee z = x_N))\urcorner$ alters the probability, so it is reasonable to believe that the latter probability is just $1 - r$, which, of course, is much smaller than r.[8] Thus we have

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

Accordingly, the conjunction

$$\begin{aligned}
&\text{prob}((\exists z)(z \in X \ \&\ \sim Tz)/B(X) \ \&\ X \\
&\quad = \{x_1, \ldots, x_N\} \ \&\ x_1, \ldots, x_N
\end{aligned}$$

$$\text{are distinct } \& \ (\forall z)(z \in X \equiv (z = x_1 \vee \ldots \vee z = x_N)) \ \& \ Tx_1 \ \& \ \ldots \ \& \ Tx_{i-1} \ \& \ Tx_{i+1} \ \& \ \ldots \ \& \ Tx_N) < r \ \& \ p_1, \ldots, p_N \text{ are distinct}$$

is warranted. Combining this with lines #3 and #4 of η_i generates a subproperty defeater for the defeasible inference from #1 to #2 in the argument η_i, as diagrammed in Figure 8. This argument is self-defeating, so the arguments for #3, #4, and #7 indirectly defeat the argument for #2, and hence they minimally defeat the argument for #2. However, the arguments for #3 and #4 are subarguments of the argument for #6, so by the definition of minimal defeat, the argument for #7 by itself minimally defeats the argument for #6. The argument

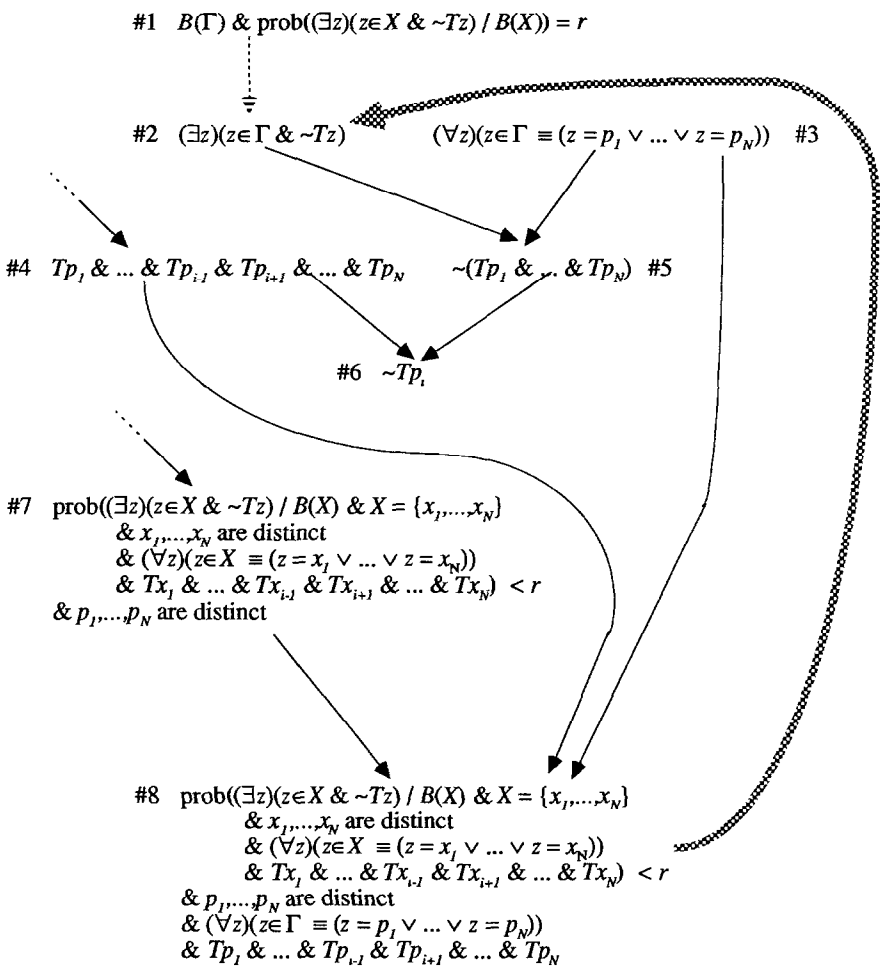


Fig. 8.

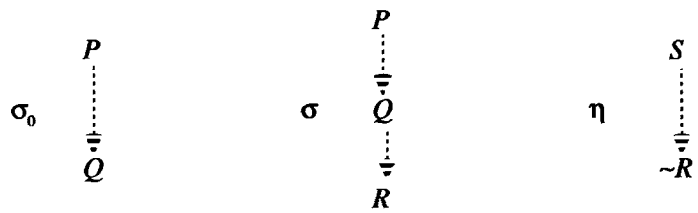
for #7 is ultimately undefeated, so it follows that the argument for #6 is defeated outright.

By hypothesis, the arguments η_i for the conclusions of the form $\sim Tp_i$ are the only arguments directly defeating the defeasible arguments for the conclusions Tp_i. Accordingly, the latter arguments are undefeated. They are in at every level. It follows that the conjunction $(Tp_1 \ \& \ \ldots \ \& \ Tp_N)$ is warranted. From this and #3 we can deduce $\sim(\exists z)(z \in \Gamma \ \& \ \sim Tz)$, so this conclusion is also warranted. Thus the defeasible argument for #2 is defeated outright, because its negation is warranted. In other words, in the paradox of the preface, we are warranted 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.[9]

What this rather complex analysis shows is 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. However, it requires an understanding of self-defeating arguments to explain why this difference makes a difference.

10. The Priority Principle

Self-defeat plays an essential role in the resolution of another potential problem for the theory of warrant. Suppose P is a prima facie reason for Q, Q is a prima facie reason for R, and S is an equally good prima facie reason for $\sim R$. Suppose P and S are both included in *input*. We can then construct the following three arguments:

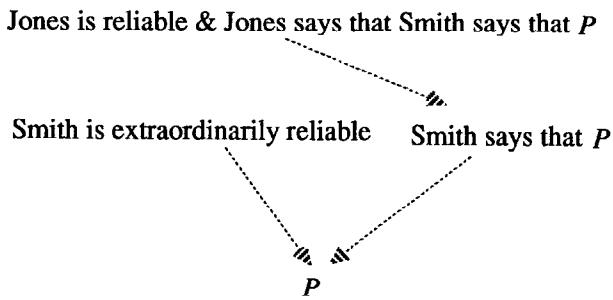


σ and η rebut one another, resulting in R undergoing provisional defeat, but σ_0 (the initial part of σ) is undefeated. This is the intuitively correct result. Given an argument like σ in which a number of prima facie reasons are strung together to obtain a final conclusion R, if we also have an equally good reason for $\sim R$ then only the last prima facie reason undergoes defeat, leaving the initial part of the argument undefeated. To illustrate this with a realistic example, perception may

give us prima facie reasons for believing of a number of individual birds that each can fly. Those conclusions jointly give us an inductive prima facie reason for believing that most birds can fly. This statistical generalization gives us a prima facie reason for believing that another bird, Tweety, can fly. Given a strong reason for believing that Tweety cannot fly (e.g., he has a broken wing), we withdraw the conclusion that Tweety can fly, taking the final prima facie reason to be defeated, but we have no inclination to let the defeat trickle back up the argument, leading us to withdraw either the statistical generalization that most birds can fly or the individual observations upon which the statistical generalization was based.

In general, given a rebutting defeater for the last step of a defeasible argument, we work backwards withdrawing conclusions until we get to the last defeasible step of the argument. Defeat must extend backwards over deductive steps, because we cannot withdraw the conclusion of a deductive step without withdrawing the premise, but when we come to a defeasible step we withdraw only the conclusion, retaining the premise and taking the prima facie reason to be defeated. I refer to this as the *priority principle* – we give priority to the earlier defeasible steps of an argument.

There are some examples that appear, at first, to be counterexamples to the priority principle. Suppose I regard Jones as of ordinary reliability, and Smith as extraordinarily reliable. Suppose Jones tells me that Smith says that P, from which I can infer defeasibly that P, but I have good reason to believe $\sim P$. Intuitively, it seems that I should not just withdraw the conclusion that P, while retaining the belief that Smith says that P. Because I regard Smith as significantly more reliable than Jones, and I must choose between disbelieving Smith and disbelieving Jones, it seems that I should disbelieve Jones and so withdraw the belief that Smith said that P. This looks like a counterexample to the priority principle. I don't think that it is, however. Consider more carefully the argument that is being defeated:

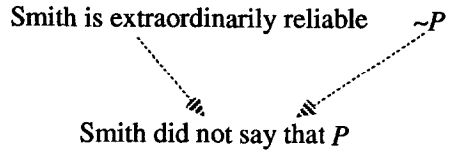


The rule of inference employed here is the statistical syllogism. However, I have argued at length in Pollock (1990) that the statistical syllogism must be augmented with an inverse rule:

THE INVERSE STATISTICAL SYLLOGISM. $\ulcorner \text{prob}(F/G) \geqslant r \mathbin{\&} \sim Fc \urcorner$ is a

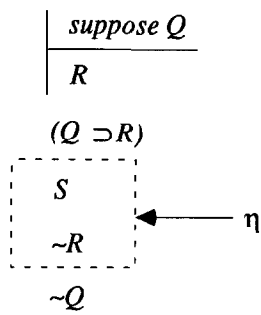
prima facie reason for $\ulcorner \sim Gc \urcorner$, the strength of the reason being a monotonic function of r.[10]

With the help of the inverse statistical syllogism, we can reason as follows:



This provides an independent source of defeat for the argument to the conclusion that Smith said that P, so that conclusion is defeated despite the priority principle. This will be true in general for arguments that proceed in terms of the statistical syllogism, because such arguments can always be contraposed, but other prima facie reasons are not similarly contrapossible, and hence this procedure will not work in general.

I allege that the priority principle is a true principle of defeasible reasoning. The *priority problem* is the problem of getting the priority principle to come out true in a theory of defeasible reasoning. The observation made above regarding the arguments σ_0, σ, and η suggests that the priority principle comes out true within the present theory of defeasible reasoning. The matter is complicated, however. The difficulty is that conditional reasoning enables us to construct more complex arguments that threaten to circumvent that reasoning and falsify the priority principle. To the above collection of arguments, let us add one more argument, μ:



μ enters into collective defeat with σ_0, with the result that unless μ is somehow defeated, σ_0 is not ultimately undefeated. But that is unreasonable – σ_0 should be undefeated. More generally, even in the absence of σ_0, an argument of the form of μ must be defeated. Otherwise, whenever we believe the denial of the conclusion of a prima facie reason, that would lead us to infer the denial of the premise, and we do not reason in that way. Notice that even if we became convinced that the priority principle is false, it is undeniable that it has some true

instances (e.g., the example given above), but the argument μ would preclude there being *any* true instances of the priority principle, and that is surely absurd.

The priority problem puzzled me for a long time, but it has a simple solution in terms of self-defeat. η defeats μ by defeating the inference of R on the supposition $\{Q\}$, and η is a subargument of μ, so μ is self-defeated. I take this to be an additional important confirmation for the account of self-defeat, because I see no way to resolve this problem without appealing to self-defeat.

This resolution of the priority problem also illustrates another important feature of the theory. In my publications on defeasible reasoning, I have vascillated between requiring that when an argument σ rebuts or undercuts an argument η, (1) the defeating line has the *same* supposition set as the defeated line, or (2) the supposition set of the defeating line is *a subset of* the supposition set of the defeated line.[11] In the present paper, I have required only containment. I will refer to this as *the subset rule*. The subset rule is essential to the preceding resolution of the priority problem. If it were required instead that the defeating line have the same supposition set as the defeated line, then η would not defeat μ. The most we could get is that μ is defeated by an argument η^* that results from reproducing η within the supposition Q:

Suppose Q
S
~R

However, this generates only collective defeat for μ – not self-defeat. Collective defeat is inadequate to solve the problem, because μ will still be in at even levels, and hence can still provisionally defeat σ_0.

The subset rule is closely related to another principle, according to which a conclusion obtained in one supposition can be automatically imported into any more inclusive supposition. This is formalized by the following inference rule:

Foreign Adoptions

If σ is an argument, $\sigma_i =$ $\langle X, p, \beta \rangle$, and $X \subseteq Y$, then $\sigma^{\frown} \langle Y, p, \langle \{i\}, \textit{foreign adoptions} \rangle \rangle$ is an argument.

Is the rule of foreign adoptions reasonable? For a while I was convinced that it was not, on the grounds that a new supposition could consist of a defeater for a previously constructed argument. Within that supposition we could not reconstruct the original argument because it would be automatically defeated. Accordingly, it seems illegitimate to simply import the conclusions of the original argument into the new supposition. For example, suppose P is a prima facie

reason for Q, Q is a prima facie reason for R, and D is a defeater for the first prima facie reason. If P is in *input*, then we can construct the following argument α:

P
Q
R

If we are automatically allowed to import the conclusions of α into arguments involving richer suppositions, then we could construct another argument β by reasoning as follows:

P

Q

R

| *Suppose D*
|---
| R
| $(D\ \&\ R)$

What is intuitively disturbing about β is that we cannot equally construct an argument β^* by reasoning as follows:

| *Suppose D*
|---
| P
| Q
| R
| $(D\ \&\ R)$

β^* is self-defeating. This seems to make β suspect, because it looks like β is just a kind of shorthand version of β^*.

Although I once found this a convincing objection to the rule of foreign adoptions, I no longer find it is compelling. We should distinguish between two kinds of 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.

These two kinds of suppositional reason appear to work in importantly different ways. In factual suppositional reasoning, because we are supposing that something *is* the case, it seems that 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. Viewing β in this light, it seems to be unobjectionable as a piece of factual suppositional reasoning, but unacceptable as a piece of counterfactual suppositional reasoning. The rule of conditionalization that is required for μ generates only material conditionals, so it seems that the relevant kind of supposition is a factual supposition rather than a counterfactual supposition. For factual suppositions, the rule of foreign adoptions seems reasonable. For exactly the same reason, the subset rule for defeat seems reasonable. That is, it seems reasonable to take one argument to defeat a second even when the relevant supposition set in the first argument is just a proper subset of the relevant supposition set in the second argument.

11. Conclusions

This completes the account of self-defeat and its impact on the analysis of warrant. I have argued that this phenomenon is of fundamental importance to the logical structure of defeasible reasoning and to understanding how to use defeasible reasoning in concrete applications. In particular, unless we can clearly differentiate between problems having the structure of the lottery paradox and problems having the structure of the paradox of the preface, any practical application of defeasible reasoning will be crippled. And I have argued that the difference between these two paradoxes turns on the phenomenon of self-defeat.

It is important to realize that this simple analysis only works subject to the pretense that all reasons are of the same strength. Otherwise we might have an argument σ supporting q and an argument η supporting $\neg q$, where σ provides a significantly better reason for q than η does for $\neg q$. In that case, rather than having collective defeat, σ should be undefeated. Space precludes my extending the present account to the case of reasons of variable strength. An earlier paper (Pollock 1991) did undertake this task. In that paper, I failed to get self-defeat right, but the present account can be incorporated into the theory of that paper without significantly altering the way in which the strengths of reasons are handled. This is because taking account of reason-strength only affects the analysis of direct defeat. The resulting modified definition of direct defeat can be plugged into the rest of the present theory without further alterations.

This theory of defeasible reasoning with self-defeat has been incorporated into a defeasible version of OSCAR. This automated reasoner engages in interest-driven suppositional reasoning, is complete for the predicate calculus, and is now a complete implementation of the theory of defeasible reasoning described in this paper. This reasoner is described more fully in Pollock (1991a).

Notes

[1] This is illustrated repeatedly in my (1974), (1986) and (1990).

[2] This characterization of warrant was presented in my (1986) and (1987). A similar proposal is contained in Horty *et al.* (1987).

[3] The lottery paradox is due to Kyburg (1961).

[4] Kyburg (1970) draws a different conclusion, namely, that we can be justified in holding inconsistent sets of beliefs, and that it is not automatically reasonable to adopt the conjunction of beliefs one justifiably holds (i.e., adjunction fails). I have argued against that view in my (1986), pp. 105–112.

[5] The paradox of the preface originated with David Makinson (1965).

[6] This is based on my discussion in (1990).

[7] See Pollock (1990) for more details on subproperty defeaters.

[8] This inference proceeds by non-classical direct inference. See Pollock (1990).

[9] If this still seems paradoxical, it is probably because one is overlooking the fact that "Books of this general sort usually contain falsehoods" formulates a *general* probability (an *indefinite probability* in the sense of Pollock, 1990), but "This book probably contains a falsehood" expresses a single-case probability (a *definite* probability). The relationship between general probabilities and single-case 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 containing a falsehood is zero. For more on direct inference, see Pollock (1990).

[10] For a more precise formulation of the inverse statistical syllogism, see Pollock (1990).

[11] In Pollock (1986, 1990, and 1990a), I required the supposition sets to be the same, but in Pollock (1990), I instead required only the subset condition. However, while the latter paper was still in press, I became convinced that that account was wrong, on the basis of the argument presented below. For the reasons presented here, I am now endorsing the subset condition again.

References

Horty, J., Thomason, R., and Touretzky, D. (1987), 'A Skeptical Theory of Inheritance in Non-Monotonic Semantic Nets', *Proceedings of AAAI-87.*

Kyburg, Henry, Jr. (1961), *Probability and the Logic of Rational Belief*, Middletown: Wesleyan University Press.

Kyburg, Henry, Jr. (1970), 'Conjunctivitis', in Marshall Swain (ed.), *Induction, Acceptance, and Rational Belief*, Dordrecht: Reidel.

Lin, Fangzhen, and Yoav Shoham (1990), 'Argument Systems: A Uniform Basis for Nonmonotonic Reasoning', Technical report No. STAN-CS-89-1243, Dept. of Computer Science, Stanford University.

Makinson, David (1965), 'The Paradox of the Preface', *Analysis* **25**, pp. 205–207.

Pollock, John (1974), *Knowledge and Justification*, Princeton.

Pollock, John (1986), *Contemporary Theories of Knowledge*, Totowa, NJ: Rowman and Littlefield.

Pollock, John (1987), 'Defeasible Reasoning', *Cognitive Science* **11**, pp. 481–518.

Pollock, John (1990), *Nomic Probability and the Foundations of Induction*, NY: Oxford University Press.

Pollock, John (1990a), *How to Build a Person*, Cambridge: Bradford/MIT Press.

Pollock, John (1991), 'A Theory of Defeasible Reasoning', *International Journal of Intelligent Systems* **6**, pp. 33–54.

Pollock, John (1991a), 'How to Reason Defeasibly', Oscar Project Technical Report.