ELSEVIER

Artificial Intelligence 67 (1994) 377–407

Artificial Intelligence

Justification and defeat

John L. Pollock*

Department of Philosophy, University of Arizona, Tucson, AZ 85721, USA

Received July 1992; revised April 1993

Abstract

This paper exhibits some problematic cases of defeasible or nonmonotonic reasoning that tend to be handled incorrectly by all of the theories of defeasible and nonmonotonic reasoning in the current literature. The paper focuses particularly on default logic, circumscription, and the author's own argument-based approach to defeasible reasoning. A proposal is made for how to deal with these problematic cases. The paper closes with a demonstration that the proposed solution is able to differentiate, in a congenial way, between cases having the structure of the lottery paradox and cases having the structure of the paradox of the preface. The algorithm proposed for computing justificational status has been implemented in the automated defeasible reasoner OSCAR.

1. Introduction

The purpose of this paper is to exhibit some problematic cases of defeasible or nonmonotonic reasoning that tend to be handled incorrectly by all of the theories of defeasible and nonmonotonic reasoning in the current literature, and then to propose a way of dealing with these cases. The paper will focus on my own argument-based approach to defeasible reasoning,[1] together with default logic[2] and circumscription.[3] The problems, however, are general ones that seem to recur within all extant theories.

I will assume a familiarity with default logic and circumscription. By contrast with either, my own argument-based approach to defeasible reasoning emphasizes that the subject is *reasoning*. For present purposes, the most convenient

* E-mail: pollock@ccit.arizona.edu.

[1] The most recent formulation of this theory is in [16–18]. It has been developed over a period of twenty-five years. See also [7–10, 12, 13, 15].

[2] See particularly [19, 20].

[3] See particularly [5, 6]

0004-3702/94/$07.00 © 1994 Elsevier Science B.V. All rights reserved
SSDI 0004-3702(93)E0039-O

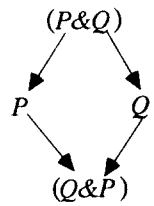

Fig. 1. An inference graph.

representation of arguments is as graphs, where the nodes represent steps of inference. I will call these *inference graphs*. A simple example is given in Fig. 1. The exact structure of the arguments and the nodes will not be important for the purposes of this paper. For instance, we can accommodate suppositional reasoning (conditionalization, reasoning by cases, reductio an absurdum, etc.) by incorporating suppositions into the nodes, but whether we do that is not germane to this paper.[4]

When a reasoner reasons, it is natural to regard it as producing a number of separate arguments aimed at supporting different conclusions. But we can also 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 can 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.

The inference relations between nodes of the inference graph are recorded in *inference links*. Where ν and η are nodes, $\langle \nu, \eta \rangle$ is an inference link iff ν was inferred (in one step) from a set of nodes one of which was η. If $\langle \nu, \eta \rangle$ is an inference link, then η is an *immediate inference ancestor* of ν. A reasoner might construct more than one argument supporting a single conclusion. It will simplify the discussion if we keep the inference links embodied in the different arguments separate. That can be done by having a different node for each argument. Nodes and their inference links then become unambiguous. This will allow us to regard the different nodes as defeated by different defeaters. 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 η to ν.

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 and how far the agent has got in its reasoning. A necessary condition for a belief to be justified 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.

[4] For more on suppositional reasoning, see [16].

Justification is defeasible in the sense that a belief may be justified at one time and become unjustified later as a result of further reasoning producing a new argument that defeats the first. This leads to a distinction between justification and warrant, where the latter is "justification in the limit", i.e., justification when all possible relevant reasoning has been completed.[5] This paper will not be concerned with warrant, but rather with the question of what beliefs are justified at any given time given the current state of the agent's inference graph.

The agent's reasoning is encoded in the inference graph, the nodes of which correspond to inferences. Defeat relations can also be encoded in the inference graph by introducing a new set of links. Where μ and ν are nodes of the inference graph, $\langle \mu, \nu \rangle$ is a *defeat link* iff ν defeats μ. These defeat relations will result in some of the nodes being defeated and others undefeated. A justified belief is one corresponding to an undefeated node of the inference graph. To complete this account, we need a characterization of the circumstances under which a node is defeated or undefeated.

The defeat status of a node is a function of its defeat relations to other nodes. The theory that will be proposed here will be independent of the exact nature of these defeat relations, but for illustrative purposes I will sketch my preferred account. According to that account, reasoning proceeds by constructing arguments, where *reasons* provide the atomic links in arguments. *Conclusive reasons* are deductive reasons, and they 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 conclusions, but it is defeasible. I will encode a reason as an ordered pair $\langle \Gamma, p \rangle$, where Γ is the set of premises of the reason and p is the conclusion. I will occasionally refer to the premises as a "reason for" the conclusion. Considerations that defeat prima facie reasons are *defeaters*. There are two importantly different kinds of defeaters. Where P is a prima facie reason for Q, R is a *rebutting defeater* iff R is a reason for denying Q. All work on nonmonotonic logic and defeasible reasoning has recognized the existence of rebutting defeaters, but it has sometimes been overlooked that there are other defeaters too.[6] 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. The preceding indicates that if P is a prima facie reason for Q, then the negation of $\ulcorner P$ wouldn't be true unless Q were true$\urcorner$ is a defeater. I will abbreviate this as $\ulcorner(P \otimes Q)\urcorner$. Such defeaters are

[5] For more on this distinction, see [12, 14, 18].

[6] The existence of defeaters other than undercutting defeaters was first pointed out in [9].

undercutting defeaters. I have argued elsewhere that rebutting defeaters and undercutting defeaters suffice for describing all defeat relations.[7] I will use rebutting defeaters and undercutting defeaters to illustrate the problems that will be discussed below, but this account of defeat relations will not be presupposed by my proposal for dealing with the problematic inferences I will describe.

Note the close parallel between prima facie reasons and defaults in default logic. A prima facie reason $\langle\{P\}, Q\rangle$ with defeaters $R_1, \ldots, R_n$ corresponds closely to a default

$$\frac{P: \sim R_1, \ldots, \sim R_n}{Q}.$$

The same relationship might be expressed in circumscription by adopting the axioms:

$$(P \& \sim ab) \supset Q\ ,$$
$$R_1 \supset ab\ .$$
$$\vdots$$
$$R_n \supset ab\ ,$$

and circumscribing ab.

2. Computing defeat status

Justified conclusions are those supported by undefeated nodes. Let us turn to the question of how the defeat status of a node of the inference graph is determined by its defeat relations to other nodes. I will begin by giving a preliminary account which captures much of the structure of the computation of defeat status, and then I will explain why it must be made more complex.

It is initially plausible that there are just two ways a node can come to be defeated. This can happen (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:

Principle 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.

[7] The argument consists of showing that large blocks of reasoning can be reconstructed using only these tools. See [10, 12, 15].

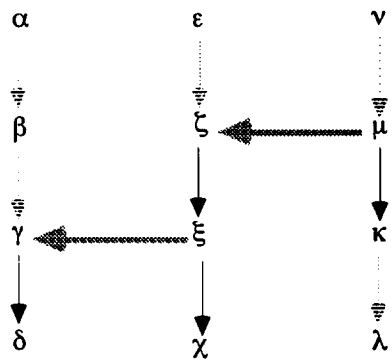

Fig. 2. An inference graph.

(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 Fig. 2, where defeasible inferences are indicated by dashed arrows, deductive inferences by solid arrows, and defeat links by arrows of the form "$\Leftarrow$". α, β, ε, ν, μ, κ, and λ are d-initial nodes, so they are undefeated. By Principle 1(3), ζ, ξ, and χ are then defeated. By Principle 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 1, but as we will see below, the different theories diverge on some complicated cases.

3. Collective and provisional defeat

I take it that Principle 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 Fig. 2, Principle 1 assigns either "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 Fig. 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 1, and then it follows that no status is assigned to any of γ, δ, ξ, or χ either.

In order to evaluate this result, we must first decide what *should* happen in cases of collective defeat. Collective defeat is familiar in AI from the discussion of *skeptical* and *credulous* reasoners (see Touretzky, Horty and Thomason [21]). Roughly, skeptical reasoners withhold belief when they have equally good reasons

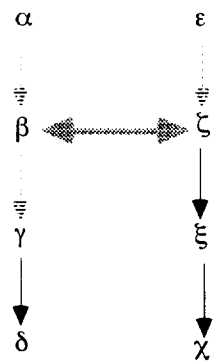

Fig. 3. Collective defeat.

for and against a conclusion, and credulous reasoners choose a conclusion at random. It has sometimes been urged that the choice between skeptical and credulous reasoners is more a matter of taste than a matter of logic, but my own view is that credulous reasoners are just wrong. Suppose you have two friends, Smith and Jones, that you regard as equally reliable. Smith approaches you in the hall and says, "It is raining outside." Jones then announces, "Don't believe him. It is a fine sunny day." If you have no other evidence regarding the weather, what should you believe? It seems obvious that you should *withhold* belief, believing neither that it is raining nor that it is not. If you were to announce, "I realize that I have no better reason for thinking that it is raining than for thinking that it is not, but I choose to believe that it is raining", no one would regard you as rational.

I have heard credulous reasoners defended on the grounds that if an agent is making practical decisions, it is better to do something rather than nothing.[8] Sometimes this seems right. For instance, if the agent is deciding where to have a picnic and the considerations favoring two sites are tied, it seems reasonable to choose at random.[9] But there are other situations in which such a policy could be disastrous. If the agent is performing medical diagnosis, and the evidence favoring two diseases is tied, we do not want the agent to decide randomly to treat the patient for one disease rather than the other. It could happen that the diseases are not serious if left untreated, but if the patient is treated for the wrong disease, that treatment will gravely exacerbate his condition. In such a case we want the agent to reserve judgment on the matter and not proceed blindly.

The difference between these two examples is that in the case of the picnic, the agent's epistemic ignorance makes the expected values of both plans for where to hold the picnic equal, and either plan is preferable to not holding the picnic at all. But in the medical diagnosis case, the agent's ignorance makes the expected value of doing nothing higher than the expected value of either plan for treatment. If

[8] Both Jon Doyle and Richmond Thomason have argued this way in conversations.
[9] This example is due to Jon Doyle (in conversation).

the agent resolved its ignorance by resolving epistemic ties at random and then acting on the basis of the conclusions thus drawn, it could not distinguish between these two kinds of cases. Instead, a rational agent should acknowledge its ignorance and take that into account in computing the expected values of plans.

The preceding considerations suggest that the controversy over skeptical and credulous reasoning stems from a confusion of epistemic reasoning (reasoning about what to believe) with practical reasoning (reasoning about what to do). In practical reasoning, if one has no basis for choosing between two alternative plans, one should choose at random. The classical illustration of this is the medieval tale of Buridan's ass who starved to death standing midway between two equally succulent bales of hay because he could not decide from which to eat. This marks an important difference between practical reasoning and epistemic reasoning. An agent making practical decisions must first decide what to believe and then use that in deciding what to do, but these are two different matters. If the evidence favoring two alternative hypotheses is equally good, the agent should record that fact and withhold belief. Subsequent practical reasoning can then decide what to do given that epistemic conclusion. In some cases it may be reasonable to choose one of the hypotheses at random and act as if it is known to be true, and in other cases more caution will be prescribed. But what must be recognized is that the design of the system of practical reasoning is a separate matter from the design of the system of epistemic reasoning that feeds information to the practical reasoner. The theory of epistemic reasoning should acknowledge ignorance rather than drawing conclusions at random. This is captured formally by the *principle of collective defeat*:

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.

Even apart from the dispute about skeptical and credulous reasoners, this way of handling collective defeat is not uncontroversial. Another example of collective defeat occurs in the lottery paradox.[10] 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 infer defeasibly that your ticket will not win.[11] But by the same reasoning, it will be reasonable to believe, for each ticket, that it will not win. These conclusions conflict jointly with something else we are warranted in believing, namely, that some ticket will win. Assuming that we cannot be warranted in believing each member of an explicitly contradictory set of propositions, it follows that we are not warranted in believing

[10] The lottery paradox is due to Kyburg [2].

[11] This reasoning proceeds in terms of the statistical syllogism. See my [15] for a book-length discussion of the statistical syllogism.

of each ticket that it will not win. This is in accordance with the principle of collective defeat. But not everyone agrees with this diagnosis of the lottery paradox. Kyburg [2] and Etherington, Kraus, and Perlis [1] both urge that it is reasonable to conclude of any individual ticket that it will not be drawn. However, there is a simple argument that seems to show that this is wrong. If I am justified in believing that ticket n will definitely not be drawn, and not just that it is improbable that ticket n will be drawn, then if I am presented with the opportunity of purchasing ticket n, it is unequivocally true that I should not purchase it. However, this conclusion would be unreasonable. No matter how improbable it is that ticket n will be drawn, if the payoff is sufficiently great, then I *should* buy the ticket. For instance, if the probability of ticket n being drawn is one in a million but the ticket costs one dollar and the payoff is one billion dollars, then rationality clearly dictates that I should buy the ticket. On the other hand, if I am justified in believing that the ticket will not be drawn, and not just that it is improbable that it will be drawn, then I am precluded from reasoning in this way. It would be irrational for me to buy the ticket, no matter what the payoff, if I am justified in believing that it will not be drawn. Accordingly, that cannot be a justified belief in the case of a fair lottery. My conclusion is that the principle of collective defeat handles the lottery paradox correctly.

This becomes an immediate criticism of a number of standard approaches to defeasible or nonmonotonic reasoning, because systems like default logic and autoepistemic logic are credulous in their standard formulations. We can, however, generate skeptical versions of these systems. For instance, we can take skeptical default logic to require that a conclusion hold in every minimal extension. Henceforth, when I refer to default logic, I will be talking about this skeptical variant of the standard theory. Circumscription is already skeptical, so no changes are required in circumscription to deal with collective defeat. There are a number of varieties of circumscription, however, and I will not be careful about distinguishing between them.[12] What I have to say about circumscription should be applicable to all the different varieties.

Collectively defeated inferences are defeated, in the sense that it is unreasonable to accept their conclusions. But principle 1 does not rule them defeated. This may be less of a problem for Principle 1 than it seems. We can regard the assignment of defeat statuses in Fig. 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 Fig. 2 with Fig. 4. In Fig. 2, ζ and ξ are defeated, and ξ thereby loses the ability to render γ defeated. In Fig. 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 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 the relationship between ζ

[12] See, for example, [3].

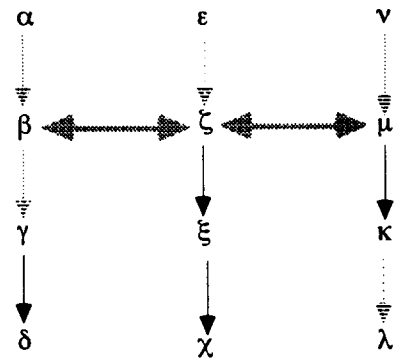

Fig. 4. The propagation of provisional defeat.

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 Fig. 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 Fig. 4 more complicated so that the collective defeat of ζ and μ involves extended reasoning, as in Fig. 5. Here, v 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

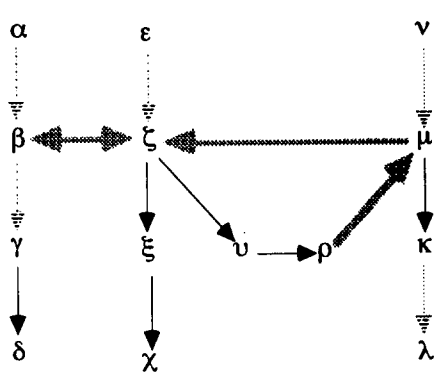

Fig. 5. The extended propagation of provisional defeat.

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 render other nodes.

The examples considered thus far can be handled by adding a fourth clause to Principle 1:

Principle 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.

This has the automatic consequence 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 2 is equivalent to the analysis of defeat I have given elsewhere.[13] However, I now believe that this account is inadequate, for the reasons that will be explored next.

4. Self-defeating arguments

The inadequacy of Principle 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 Fig. 6. The nodes supporting R and $\sim R$ (henceforth $\langle\langle R\rangle\rangle$ and $\langle\langle\sim R\rangle\rangle$)

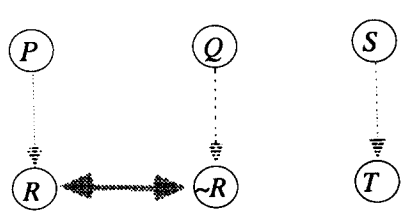

Fig. 6. A simple inference graph.

[13] First in [11], and then in my [12, 13].

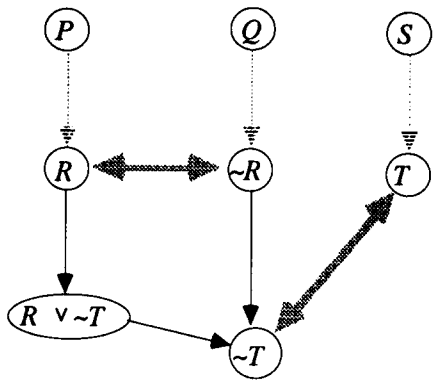

Fig. 7. Self-defeat.

collectively defeat one another, but $\langle\langle T \rangle\rangle$ should be independent of either and undefeated. The difficulty is that we can extend the inference graph as in Fig. 7. Here I have used a standard strategy for deriving an arbitrary conclusion from a contradiction. The problem is now that $\langle\langle \sim T \rangle\rangle$ rebuts $\langle\langle T \rangle\rangle$. According to Principle 2, $\langle\langle \sim R \rangle\rangle$, and hence $\langle\langle \sim T \rangle\rangle$, are provisionally defeated, but then it follows that $\langle\langle T \rangle\rangle$ 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. What this example shows is that nodes inferred from provisionally defeated nodes are not always provisionally defeated. In Fig. 7, $\langle\langle \sim T \rangle\rangle$ must be defeated outright. There is no way to get this result from Principle 2. My diagnosis of the difficulty is that the argument supporting $\langle\langle \sim T \rangle\rangle$ is "internally defective". It is *self-defeating* in the sense that some of its steps are defeaters for others. By Principle 2, this means that those inferences enter into collective defeat with one another, and hence $\langle\langle \sim T \rangle\rangle$ is provisionally defeated, but my suggestion is that this should be regarded as a more serious defect—one which leaves $\langle\langle \sim T \rangle\rangle$ 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:

Definition 1. A node η is *self-defeating* iff some of its inclusive inference ancestors defeat others.

Principle 2 should be modified so that self-defeating nodes are defeated outright rather than just provisionally.

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

However, 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 one 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 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 $\sim R$ by noting that the $\sim T_i$ jointly entail $\sim R$. This is because if none of the tickets is drawn then the lottery is not fair. This is diagrammed in Fig. 8. The difficulty is now that $\langle\langle \sim R \rangle\rangle$ rebuts $\langle\langle R \rangle\rangle$. Thus by Principle 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. Obviously, 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 $\langle\langle \sim R \rangle\rangle$ is self-defeating.

Default logic gets the example of Fig. 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 Figs. 7 and 8 can be avoided by ruling that self-defeating nodes are defeated outright—not just provisionally. As they are defeated outright, they cannot enter into collective defeat with other nodes, and so the nodes $\langle\langle \sim R \rangle\rangle$ and $\langle\langle \sim T \rangle\rangle$ in the preceding

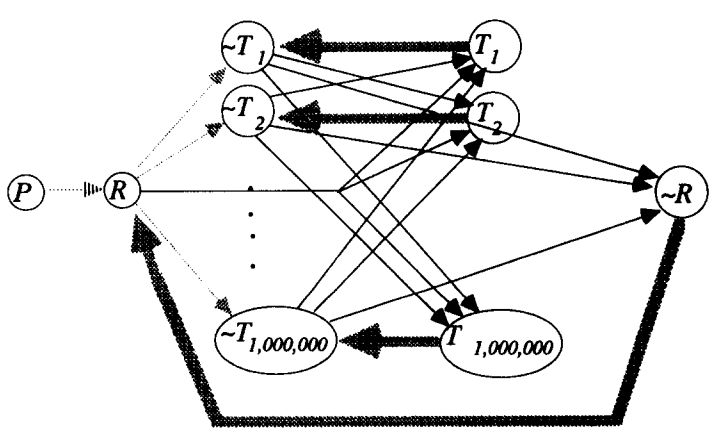

Fig. 8. The lottery paradox paradox.

two examples are defeated outright, as they should be. This can be accomplished by revising Principle 2 as follows:

Principle 3.
(1) D-initials 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 [17]. However, as will be seen, 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 Fig. 9. The node $\langle\langle P \otimes Q \rangle\rangle$ is self-defeating, because it defeats one of its own ancestors. Thus by Principle 3, it is defeated outright. It then follows from Principle 3 that the remaining nodes are undefeated. But this is most peculiar, because $\langle\langle P \otimes Q \rangle\rangle$ is a deductive consequence of $\langle\langle R \rangle\rangle$. If a node is undefeated, its deductive consequences should also be undefeated. Conversely, if 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 $\langle\langle R \rangle\rangle$ should be defeated. What about $\langle\langle Q \rangle\rangle$? 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

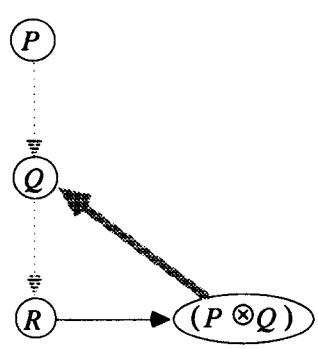

Fig. 9. No nearest defeasible ancestor is defeated.

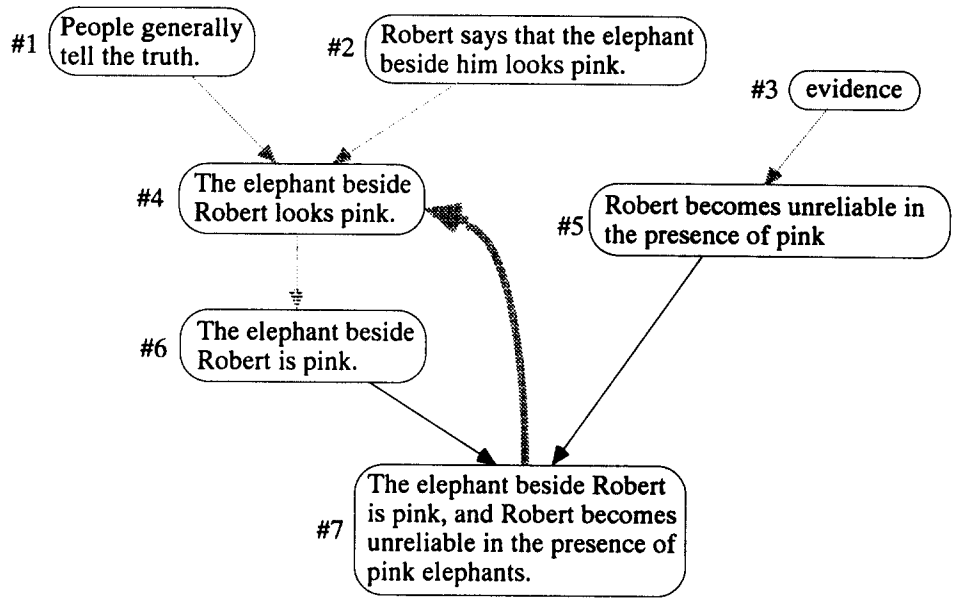

Fig. 10. Argument with a self-defeating conclusion.

elephants, gives us a defeater for our reason for thinking that the elephant looks pink. These relations can be diagrammed as in Fig. 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 #6 should be defeated *by* having #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 $\langle\langle Q \rangle\rangle$ as defeated in Fig. 9. Neither of these conclusions is forthcoming from Principle 3. 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.

It turns out that circumscription gives the right result in Figs. 9 and 10. In Fig. 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 Fig. 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 counterexample to default logic.

5. A new approach

Summing up the previous discussion, we find that default logic and circumscription handle some of the problem cases correctly and some incorrectly. The cases

in which they fail tend to be cases in which they are not sufficiently sensitive to the structure of the arguments For example, in Fig. 8, circumscription gets self-defeat wrong, and in Fig. 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 2 and 3 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—Fig. 7. Default logic and circumscription get this example right, but Principle 2 gets it wrong, necessitating the explicit appeal to self-defeat in Principle 3. 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 2, on the other hand, washes these relationships out, just assigning a blanket status of "provisionally defeated" to all provisionally defeated propositions.

The conclusion I want to draw from this 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 1 and 2. Let us define a *status assignment* to be an assignment of defeat status that is consistent with the rules of Principle 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 will assign 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:

Definition 2. 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 an immediate ancestor that is

assigned "defeated", or there is a node β that defeats α and is assigned "undefeated".

The proposal is then:

Principle 4. 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 Fig. 7, there is one status assignment assigning "defeated" to $\langle\langle R \rangle\rangle$ and "undefeated" to $\langle\langle \sim R \rangle\rangle$, and another status assignment assigning the opposite statuses. On both assignments, $\langle\langle \sim T \rangle\rangle$ is assigned "defeated", so by Principle 4, $\langle\langle \sim T \rangle\rangle$ is defeated outright. Fig. 8 is analogous. In Fig. 8, for each i there is a status assignment assigning "defeated" to $\langle\langle \sim T \rangle\rangle$ but assigning "undefeated" to $\langle\langle R \rangle\rangle$ and all the other $\langle\langle \sim T \rangle\rangle$'s. Every such status assignment assigns "defeated" to $\langle\langle \sim R \rangle\rangle$. Thus by Principle 4, $\langle\langle R \rangle\rangle$ is undefeated and $\langle\langle \sim R \rangle\rangle$ is defeated outright, while all of the $\langle\langle \sim T \rangle\rangle$'s are provisionally defeated.

Of the examples from the first part of the paper, Principle 4 is able to handle all but that of Fig. 9. In Fig. 9, something unexpected happens. Any status assignment assigning "undefeated" to $\langle\langle Q \rangle\rangle$ will also assign "undefeated" to $\langle\langle R \rangle\rangle$ and to $\langle\langle P \otimes Q \rangle\rangle$, but then it must instead assign "defeated" to $\langle\langle Q \rangle\rangle$. Thus no status assignment can assign "undefeated" to $\langle\langle Q \rangle\rangle$. However, no status assignment can assign "defeated" to $\langle\langle Q \rangle\rangle$ either, because then it would have to assign "defeated" to $\langle\langle R \rangle\rangle$ and $\langle\langle P \otimes Q \rangle\rangle$ as well, from which it follows that it must instead assign "undefeated" to $\langle\langle Q \rangle\rangle$. What this shows is that no status assignments are possible for the inference graph of Fig. 9. We can construct other examples of this same phenomenon. The simplest involve odd-length defeat cycles. Consider the inference graph diagrammed in Fig. 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 let A be "Smith says that it is raining" and B be "It is raining". Intuitively, $\langle\langle Q \rangle\rangle$, $\langle\langle S \rangle\rangle$, and $\langle\langle U \rangle\rangle$ ought to collectively defeat one another, and then because $\langle\langle Q \rangle\rangle$ is provisionally defeated, $\langle\langle B \rangle\rangle$ should be provisionally defeated. That is precisely

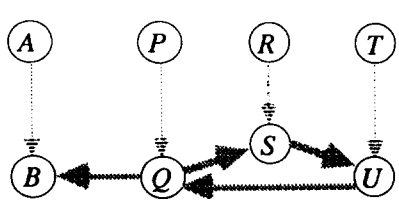

Fig. 11. A three-membered defeat cycle.

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 1. For example, if $\langle\langle Q \rangle\rangle$ is assigned "undefeated", $\langle\langle S \rangle\rangle$ must be assigned "defeated", and then $\langle\langle U \rangle\rangle$ must be assigned "undefeated", with the result that $\langle\langle Q \rangle\rangle$ 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 Fig. 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 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:

Definition 3. 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 defeating α are assigned "defeated"; and
(3) σ assigns "defeated" to a node α iff either α has an 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:

Definition 4. 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 4 handles the examples of Figs. 9 and 11 properly. In Fig. 9, nodes $\langle\langle Q \rangle\rangle$, $\langle\langle R \rangle\rangle$, and $\langle\langle P \otimes Q \rangle\rangle$ turn out to be defeated, and in Fig. 11, nodes $\langle\langle B \rangle\rangle$, $\langle\langle Q \rangle\rangle$, $\langle\langle S \rangle\rangle$, and $\langle\langle U \rangle\rangle$ are defeated. This is my final proposal for the analysis of defeat for nodes of the inference graph.[14]

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

(1) A node α is undefeated iff all immediate ancestors of α are undefeated and all nodes defeating α are defeated outright.
(2) If some immediate ancestor of α is defeated outright, then α is defeated outright.

[14] In a number of earlier publications [11–13, 15, 16], I proposed that defeat could be 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 feel that obscured the proper treatment of self-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).

(3) If some immediate ancestor of α is provisionally defeated, then α is either provisionally defeated or defeated outright.
(4) If some node defeating α is undefeated, then α is defeated outright.
(5) If α is self-defeating, then α is defeated outright.

6. The paradox of the preface[15]

Much of my work on the analysis of defeat has been driven by an 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 have discussed the lottery paradox above, and maintained that it can be regarded as a straightforward case of collective defeat. Contrast that with the paradox of the preface (due to David Makinson [4]), which can be presented as follows (see [17]):

> 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 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

[15] This section is based upon [17], but concludes by giving a different diagnosis of the paradox—one based upon the new analysis constituted by Principle 4.

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 as 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.

What will be shown now is that the paradox of the preface can be resolved by appealing to the analysis of defeat proposed above.[16] 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). 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:

$$\text{prob}(Tz \mid z \in X \,\&\, B(X)) = r\,,$$

$$\text{prob}((\exists z)(z \in X \,\&\, {\sim}Tz) \mid B(X)) \geqslant r\,.$$

The latter 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 scheme diagrammed in Fig. 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 #1

[16] This corrects the discussion in my [15, 17].

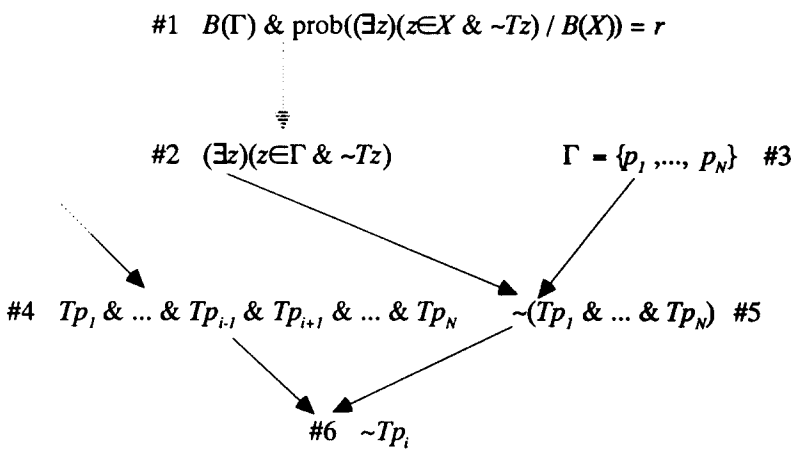

Fig. 12. The paradox of the preface.

to #2 arises from establishing anything of the following form (for any property C):

$$C(\Gamma)\,\&\,\mathrm{prob}((\exists z)(z \in X \,\&\, \sim Tz) \mid B(X) \,\&\, C(X)) < r\,.^{17}$$

It is shown in [15, p. 251] that

$$\begin{aligned}\mathrm{prob}((\exists z)(z \in X \,\&\, \sim Tz) \mid B(X) \,\&\, X = \{x_1, \ldots, x_N\} \,\&\\ x_1, \ldots, x_N \text{ are distinct}^{18} \,\&\\ Tx_1 \,\&\, \cdots \,\&\, Tx_{i-1} \,\&\, Tx_{i+1} \,\&\, \cdots \,\&\, Tx_N)\\ = \mathrm{prob}(\sim Tx_i \mid B(X) \,\&\, X = \{x_1, \ldots, x_N\} \,\&\, x_1, \ldots, x_N \text{ are distinct } \&\\ Tx_1 \,\&\, \cdots \,\&\, Tx_{i-1} \,\&\, Tx_{i+1} \,\&\, \cdots \,\&\, Tx_N)\,.\end{aligned}$$

Now we come to the point at which 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,

$$\begin{aligned}\mathrm{prob}(\sim Tx_i \mid B(X) \,\&\, X = \{x_1, \ldots, x_N\} \,\&\, x_1, \ldots, x_N \text{ are distinct } \&\\ Tx_1 \,\&\, \cdots \,\&\, Tx_{i-1} \,\&\, Tx_{i+1} \,\&\, \cdots \,\&\, Tx_N)\\ \leqslant \mathrm{prob}(\sim Tx_i \mid B(X) \,\&\, X = \{x_1, \ldots, x_N\} \,\&\, x_1, \ldots, x_N \text{ are distinct}).\end{aligned}$$

In other words, the different claims in Γ are not negatively relevant to one another. For example, the 10,000 color judgments were assumed to be in-

[17] See [15] for more details on subproperty defeaters.

[18] "$x_1, \ldots, x_n$ are distinct" means "$x_1, \ldots, x_n$ are n different objects".

dependent 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 $\ulcorner X = \{x_1, \ldots, x_N\} \,\&\, x_1, \ldots, x_N$ 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.[19] Thus we have

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

Accordingly, the conjunction

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

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

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. Fig. 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 Fig. 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_2, and Q_3. This assigns "undefeated" to the nodes supporting S_1, S_2, and S_3, and "defeated" to the nodes supported ${\sim}Q_1$, ${\sim}Q_2$, ${\sim}Q_3$, 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 undefeated, 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". This has the result that the node supporting ${\sim}Q_i$ must be assigned

[19] This inference proceeds by non-classical direct inference. See [15].

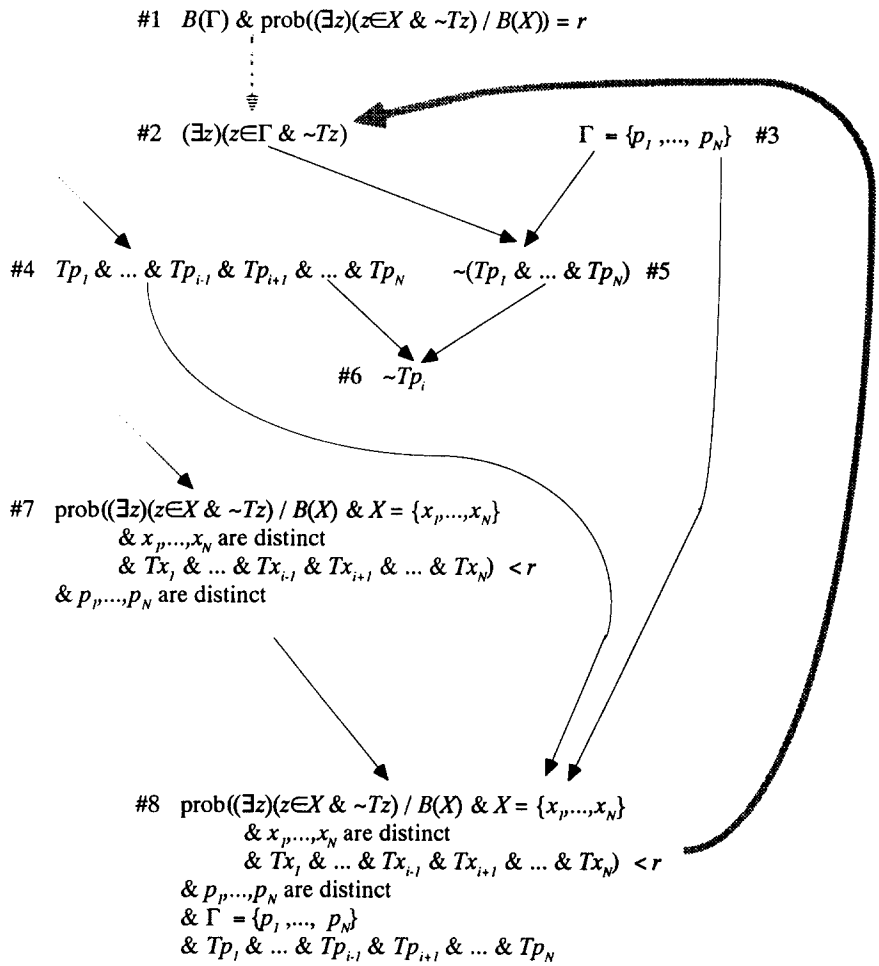

Fig. 13. Resolution of the paradox of the preface.

"defeated". That is the only node defeating that supporting Q_i, 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 Fig. 13. The nodes supporting the

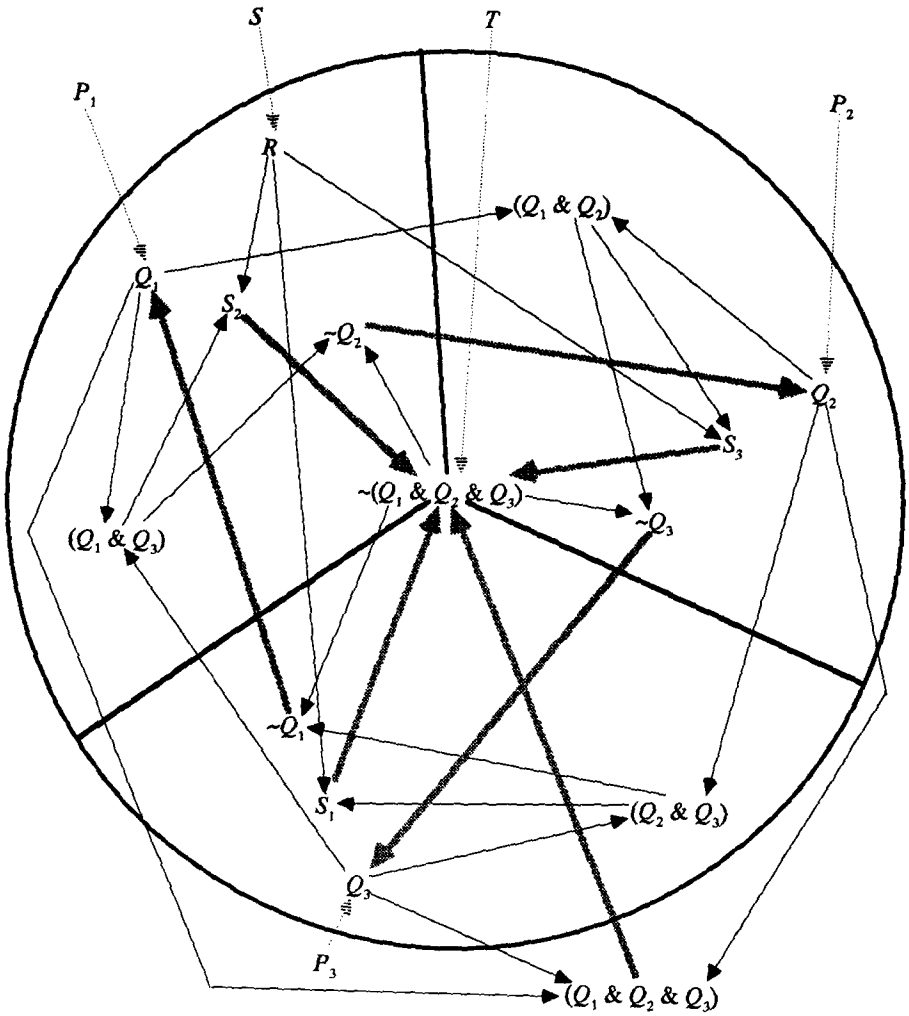

Fig. 14. The structure of the paradox of the preface.

$\sim Q_i$'s correspond to node #6, the node supporting $\sim(Q_1 \& Q_2 \& Q_3)$ corresponds to nodes #2 and #5, and the 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 Fig. 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 $(Tp_1 \& \cdots \& Tp_N)$ is justified. In 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.[20]

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. It is the analysis of defeat embodied in Principle 4 that enables us to draw these congenial conclusions.

7. Implementation

The automated defeasible reasoner OSCAR was described in [18]. The present version of OSCAR uses Principle 4 for the computation of what beliefs are justified. A direct implementation of Principle 4 would have us look at all possible partial assignments of "defeated" and "undefeated" to the nodes of the inference graph, determine which are status assignments by checking them for consistency and maximality (which will eliminate most of them), and then compute defeat status on that basis. That approach is combinatorially impossible for large inference-graphs. For an inference graph with n nodes, we would have to generate and check 3^n assignments. OSCAR employs a more efficient algorithm for generating the status assignments used for computing defeat status.

For purposes of implementation, it is convenient to regard partial status assignments as three-valued functions, assigning either "defeated", "undefeated", or "unassigned" to every member of the inference graph. Given an assignment to a subset of the nodes of an inference-graph, define the *assignment-closure* of that assignment to be the assignment that results from recursively applying the following three rules until no further nodes receive values or some node receives inconsistent values (i.e., it is assigned both "defeated" and "undefeated"):

- if all members of the basis of a node have been assigned "undefeated" and all defeaters for the node have been assigned "defeated", assign "undefeated" to the node;
- if some member of the basis of a node has been assigned "defeated" or some

[20] 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 containing a falsehood is zero. For more on direct inference, see [15].

defeater for the node has been assigned "undefeated", assign "defeated" to the node;

- if all members of the basis of a node and the node-defeaters of a node have received assignments, no member of the node-basis has been assigned "defeated", no member of the node-defeaters has been assigned "undefeated", and some member of either the node-basis or node-defeaters has been assigned "unassigned", then assign "unassigned" to the node;
- if some node is assigned two different values, the assignment-closure is empty.

If we did not have to worry about the fact that some maximal assignments may be partial assignments, a reasonably efficient algorithm for generating status assignments for inference graphs would be as follows:

Algorithm 1.

(1) Let σ_0 be the assignment-closure of the partial assignment that assigns "undefeated" to all initial nodes, and let *P-ass* $= \{\sigma_0\}$.
(2) Let *Ass* $= \emptyset$.
(3) Repeat (a)–(c) until no new assignments are generated:
  (a) If *P-ass* is empty, exit the loop.
  (b) Let σ be the first member of *P-ass*:
    - Delete σ from *P-ass*.
    - Let n be a node which has not been assigned a status but for which all members of the node-basis have been assigned "undefeated".
    - If there is no such node as n, insert $\langle \sigma, A \rangle$ into *Ass*.
    - If there is such an n then:
      - Let S be the set of all assignments that result from extending σ by assigning "defeated" and "undefeated" to n.
      - Insert all non-empty assignment-closures of members of S into *P-ass*.
  (c) If *Ass* is unchanged, exit the loop.
(4) Return *Ass* as the set of assignments.

This algorithm generates all total assignments by building them up recursively from below (ordering nodes in terms of the "inference–ancestor" relation). When this leaves the status of a node undetermined, the algorithm considers all possible ways of assigning values to that node, and later removes any combinations of such "arbitrary" assignments whose assignment-closures prove to be inconsistent. The general idea is that assignments are generated recursively insofar as possible, but when that is not possible a generate-and-test procedure is used.

To modify the above algorithm so that it will generate all maximal partial assignments, instead of just deleting inconsistent arbitrary assignments, we must look at proper sub-assignments of them. When such a proper sub-assignment has a consistent assignment-closure, and it is not a proper sub-assignment of any other consistent assignment, then it must be included among the maximal partial

assignments. To manage this, the algorithm must keep track of what arbitrary assignments have been made in the course of constructing an assignment. Let σ_0 be the assignment-closure of the partial assignment that assigns "undefeated" to all initials nodes. Let us take an "annotated assignment" to be a pair $\langle \sigma, A \rangle$ where A is a set of arbitrary assignments and σ is the assignment-closure of $\sigma_0 \cup A$.

Algorithm. COMPUTE-ASSIGNMENTS.

(1) Let σ_0 be the assignment-closure of the partial assignment that assigns "undefeated" to all initial nodes, and let *P-ass* $= \{\langle \sigma_0, \emptyset \rangle\}$.
(2) Let *Ass* $= \emptyset$.
(3) Repeat (a) and (b) until an exit instruction is encountered:
  (a) If *P-ass* is empty, exit the loop.
  (b) Let $\langle \sigma, A \rangle$ be the first member of *P-ass*;
    - Delete $\langle \sigma, A \rangle$ from *P-ass*.
    - Let n be a node which has not been assigned a status but for which all members of the node-basis have been assigned "undefeated".
    - If there is no such node as n, insert $\langle \sigma, A \rangle$ into *Ass*.
    - If there is such an n then:
      - Let *Ass** be the set of all $A \cup X^*$ such that X^* is an arbitrary assignment of "defeated" or "undefeated" to n.
      - Let S be the set of all maximal sub-assignments S^* of members of *Ass** such that the assignment-closure of $S^* \cup \sigma_0$ is non-empty.
      - For each member A of S^*:
        - If any member of *P-ass* is a sub-assignment of A, delete it from *P-ass*.
        - If any member of *Ass* is a sub-assignment of A, delete it from *Ass*.
        - If A is not a sub-assignment of any member of *P-ass* or *Ass*, insert A into *P-ass*.
(4) Return as the set of assignments the set of all σ such that for some A, $\langle \sigma, A \rangle$ is in *Ass*.

The correctness of this algorithm turns on the following observations:

(1) Every partial assignment can be generated as the assignment-closure of the assignment to the initial nodes and an arbitrary assignment to some otherwise undetermined nodes.
(2) If a partial assignment is inconsistent, so is every extension of it.

The algorithm makes use of (1) in the same way the previous algorithm did. In light of (2) in ruling out inconsistent assignments, we can shrink the search space by ruling out inconsistent sub-assignments and then only test for consistency the extensions of the remaining consistent sub-assignments.

To illustrate the algorithm, the following is a trace of its application to Fig. 14 (the paradox of the preface):

```
Problem #14
Base assignment:
((R undefeated) (P1 undefeated) (P2 undefeated) (P3 undefeated) (S undefeated) (T undefeated))
= = = new loop = = =
Contents of P-ass:
(
((R undefeated) P1 undefeated) (P2 undefeated) (P3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ()
)
-------------
Extending the following partial assignment:
(
((R undefeated) (P1 undefeated) (P2 undefeated) (P3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ()
)
Extending assignment by looking at node ~(Q1 & Q2 & Q3)
Maximal consistent assignments to new node:
(
((S3 undefeated) (S2 undefeated) (S1 undefeated) ((Q2 & Q3) undefeated) ((Q1 & Q3) undefeated)
 ((Q1 & Q2) undefeated) (Q1 undefeated) (Q2 undefeated) (Q3 undefeated) (~Q3 defeated)
 (~Q2 defeated) (~Q1 defeated) (~(Q1 & Q2 & Q3) defeated) (R undefeated) (P1 undefeated)
 (P2 undefeated) (P3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) defeated))
)
(
((~(Q1 & Q2 & Q3) undefeated) (R undefeated) (P1 undefeated) (P2 undefeated) (P3 undefeated)
 (S undefeated) (T undefeated))
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) undefeated))
)

Considering:
(
((S3 undefeated) (S2 undefeated) (S1 undefeated) ((Q2 & Q3) undefeated) ((Q1 & Q3) undefeated)
 ((Q1 & Q2) undefeated) (Q1 undefeated) (Q2 undefeated) (Q3 undefeated) (~Q3 defeated)
 (~Q2 defeated) (~Q1 defeated) (~(Q1 & Q2 & Q3) defeated) (R undefeated) (P1 undefeated)
 (P2 undefeated) (P3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) defeated))
)

Adding to P-ass: (
((S3 undefeated) (S2 undefeated) (S1 undefeated) ((Q2 & Q3) undefeated) ((Q1 & Q3) undefeated)
 ((Q1 & Q2) undefeated) (Q1 undefeated) (Q2 undefeated) (Q3 undefeated) (~Q3 defeated)
 (~Q2 defeated) (~Q1 defeated) (~(Q1 & Q2 & Q3) defeated) (R undefeated) (P1 undefeated)
 (P2 undefeated) (P3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) defeated))
)

Considering:
(
((~(Q1 & Q2 & Q3) undefeated) (R undefeated) (P1 undefeated) (P2 undefeated) (P3 undefeated)
(S undefeated) (T undefeated))
```

```
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) undefeated))
)

Adding to P-ass: (
((~(Q1 & Q2 & Q3) undefeated) (R undefeated) (P1 undefeated) (P2 undefeated) (P3 undefeated)
 (S undefeated) (T undefeated))
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) undefeated))
)

= = = new loop = = =
Contents of P-ass:
(
((~(Q1 & Q2 & Q3) undefeated) (R undefeated) (P1 undefeated) (P2 undefeated) (P3 undefeated)
 (S undefeated) (T undefeated))
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) undefeated))
)
(
((S3 undefeated) (S2 undefeated) (S1 undefeated) ((Q2 & Q3) undefeated) ((Q1 & Q3) undefeated)
 ((Q1 & Q2) undefeated) (Q1 undefeated) (Q2 undefeated) (Q3 undefeated) (~Q3 defeated)
 (~Q2 defeated) (~Q1 defeated) (~(Q1 & Q2 & Q3) defeated) (R undefeated) (P1 undefeated)
 (P2 undefeated) (P3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) defeated))
)

-------------
Extending the following partial assignment:
(
((~(Q1 & Q2 & Q3) undefeated) (R undefeated) (P1 undefeated) (P2 undefeated) (P3 undefeated)
 (S undefeated) (T undefeated))
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) undefeated))
)

Extending assignment by looking at node Q3
Maximal consistent assignments to new node:
(
((Q3 undefeated) (~(Q1 & Q2 & Q3) undefeated) (R undefeated) (P1 undefeated) (P2 undefeated)
 (P3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ((Q3 undefeated) (~(Q1 & Q2 & Q3) undefeated))
)

Considering:
(
((Q3 undefeated) (~(Q1 & Q2 & Q3) undefeated) (R undefeated) (P1 undefeated) (P2 undefeated)
 (P3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ((Q3 undefeated) (~(Q1 & Q2 & Q3) undefeated))
)

Adding to P-ass: (
((Q3 undefeated) (~(Q1 & Q2 & Q3) undefeated) (R undefeated) (P1 undefeated) (P2 undefeated)
 (P3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ((Q3 undefeated) (~(Q1 & Q2 & Q3) undefeated))
)

= = = new loop = = =
Contents of P-ass:
(
```

((Q3 undefeated) (~(Q1 & Q2 & Q3) undefeated) (R undefeated) (P1 undefeated) (P2 undefeated)
(P_3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ((Q3 undefeated) (~(Q1 & Q2 & Q3) undefeated))
)
(
((S3 undefeated) (S2 undefeated) (S1 undefeated) ((Q2 & Q3) undefeated) ((Q1 & Q3) undefeated)
((Q1 & Q2) undefeated) (Q1 undefeated) (Q2 undefeated) (Q3 undefeated) (~Q3 defeated)
(~Q2 defeated) (~Q1 defeated) (~(Q1 & Q2 & Q3) defeated) (R undefeated) (P1 undefeated)
(P2 undefeated) (P3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) defeated))
)

Extending the following partial assignment:
(
((Q3 undefeated) (~(Q1 & Q2 & Q3) undefeated) (R undefeated) (P1 undefeated) (P2 undefeated)
(P3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ((Q3 undefeated) (~(Q1 & Q2 & Q3) undefeated))
)

Extending assignment by looking at node Q2
Maximal consistent assignment to new node:
(
((S1 undefeated) ((Q2 & Q3) undefeated) (~(Q1 & Q2 & Q3) defeated (Q2 undefeated)
(Q3 undefeated) (R undefeated) (P1 undefeated) (P2 undefeated) (P3 undefeated)
(S undefeated) (T undefeated))
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) defeated) (Q2 undefeated) (Q3 undefeated))
)

Considering:
(
((S1 undefeated) ((Q2 & Q3) undefeated) (~(Q1 & Q2 & Q3) defeated) (Q2 undefeated)
(Q3 undefeated) R undefeated) (P1 undefeated) (P2 undefeated) (P3 undefeated)
(S undefeated) (T undefeated))
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) defeated) (Q2 undefeated) (Q3 undefeated))
)
This is a sub-assignment of a member of P-Ass.

= = = new loop = = =
Contents of P-ass:
(
((S3 undefeated) (S2 undefeated) (S1 undefeated) ((Q2 & Q3) undefeated) ((Q1 & Q3) undefeated)
((Q1 & Q2) undefeated) (Q1 undefeated) (Q2 undefeated) (Q3 undefeated) (~Q3 defeated)
(~Q2 defeated) (~Q1 defeated) (~(Q1 & Q2 & Q3) defeated (R undefeated) (P1 undefeated)
(P2 undefeated) (P3 undefeated) (S undefeated) (T undefeated))
arbitrary part of assignment: ((~(Q1 & Q2 & Q3) defeated))
)

Assignment found:
((S3 undefeated) (S2 undefeated) (S1 undefeated) ((Q2 & Q3) undefeated) ((Q1 & Q3) undefeated)
((Q1 & Q2) undefeated) (Q1 undefeated) (Q2 undefeated) (Q3 undefeated) (~Q3 defeated)
(~Q2 defeated) (~Q1 defeated) (~(Q1 & Q2 & Q3) defeated) (R undefeated) (P1 undefeated)
(P2 undefeated) (P3 undefeated) (S undefeated) (T undefeated))

1 assignment

The algorithm employed by the present version of OSCAR is a refinement of the preceding algorithm.

8. Conclusions

The main conclusion of this paper is that familiar theories of defeasible and nonmonotonic reasoning fail to deal adequately with some kinds of complicated argument structures. This has been demonstrated by looking at default logic, circumscription, and my own argument-based theory of defeasible reasoning. What is required is a new analysis of defeat that combines insights from both the argument-based approach and default logic and circumscription. Such an analysis is proposed in Principle 4. It was shown that, among other things, this new analysis is able to discriminate between cases having the structure of the lottery paradox and cases having the structure of the paradox of the preface. The defeat algorithm described by Principle 4 has been incorporated into the general defeasible reasoner OSCAR, with the result that OSCAR is able to reason correctly in connection with all of the examples discussed in this paper. (OSCAR can be obtained at no cost from the author.)

References

[1] D.W. Etherington, S. Kraus and D. Perlis, Nonmonotonicity and the scope of reasoning, *Artif. Intell.* **52** (1991) 221–261.
[2] H.E. Kyburg Jr, *Probability and the Logic of Rational Belief* (Wesleyan University Press, Middletown, CT, 1961).
[3] V. Lifschitz Pointwise circumscription, in: M. Ginsberg, ed., *Readings in Nonmonotonic Reasoning* (Morgan Kaufmann, Los Altos, CA, 1987) 179–193.
[4] D. Makinson, The paradox of the preface, *Analysis* **25** (1965) 205–207.
[5] J. McCarthy, Circumscription—a form of non-monotonic reasoning, *Artif. Intell.* **13** (1980) 27–39.
[6] J. McCarthy, Applications of circumscription to formalizing common-sense knowledge, in: *Proceedings Workshop on Nonmonotonic Reasoning*, New Paltz, NY (1984); also: *Artif. Intell.* **28** (1986) 89–116.
[7] J.L. Pollock, Analyticity and implication, Ph.D. Dissertation, University of California, Berkeley, CA (1965).
[8] J.L. Pollock, Criteria and our knowledge of the material world, *Philos. Rev.* **76** (1967) 28–60.
[9] J.L. Pollock, The structure of epistemic justification, *Am. Philos. Q.* **4** (1970) 62–78.
[10] J.L. Pollock, *Knowledge and Justification* (Princeton University Press, Princeton, NJ, 1974).
[11] J.L. Pollock, Reasons and reasoning, in: *Proceedings American Philosophical Association Symposium*, Denver, CO (1979).
[12] J.L. Pollock, *Contemporary Theories of Knowledge* (Rowman and Littlefield, Totowa, NJ, 1986).
[13] J.L. Pollock, Defeasible reasoning, *Cogn. Sci.* **11** (1987) 481–518.
[14] J.L. Pollock, *How to Build a Person; a Prolegomenon* (Bradford/MIT Press, Cambridge, MA, 1989).
[15] J.L. Pollock, *Nomic Probability and the Foundations of Induction* (Oxford University Press, New York, 1990).
[16] J.L. Pollock, A theory of defeasible reasoning, *Int. J. Intell. Syst.* **6** (1990) 33–54.

[17] J.L. Pollock, Self-defeating arguments, *Minds and Machines* **1** (1991) 367–392.
[18] J.L. Pollock, How to reason defeasibly, *Artif. Intell.* **57** (1992) 1–42.
[19] R. Reiter, A logic for default reasoning, *Artif. Intell.* **13** (1980) 81–132.
[20] R. Reiter and G. Criscuolo, On interacting defaults, in: *Proceedings IJCAI-81*, Vancouver, BC (1981) 270–276.
[21] D.S. Touretzky, J.F. Horty, and R.H. Thomason, A clash of intuitions: the current state of nonmonotonic multiple inheritance systems, in: *Proceedings IJCAI-87*, Milan, Italy (1987) 476–482.