A New Semantics for Defeasible Reasoning

John L. Pollock
Department of Philosophy
University of Arizona
Tucson, Arizona 85721
pollock@arizona.edu
http://www.u.arizona.edu/~pollock

1. Defeasible Reasoning

The ultimate aspiration of AI is that of building agents with human-level intelligence capable of functioning in environments of real-world complexity. I refer to these as GIAs — generally intelligent agents. Human beings are stereotypical GIAs. OSCAR is a cognitive architecture for GIAs based on a general theory of defeasible reasoning (references). This chapter describes a newly developed semantics for the OSCAR system of defeasible reasoning.

One of the most important constraints on GIAs is that they must be able to function in environments in which they have relatively little knowledge. Reflect on the fact that you are a GIA, and consider how little you really know about the world. Our knowledge of individual matters of fact is worse than just gappy — it is sparse. Consider what you know about individual grains of sand, individual cats in China, or for that matter, most individual people. Our knowledge of general matters of fact is equally sparse. What do you know about where the fish are apt to be biting in Piña Blanca Lake, how the Brazilians will vote in their next general election, how flavors are generated in cooking, or myriad other general facts that might turn out to be relevant to your life? The proportion of all the facts about the world that are known to any one person is miniscule. Nevertheless, people, and GIAs in general, must reason about the world and make decisions about how to act. How can they do that?

Generally, we reason on the basis of what we do know, and assume that what we do not know would not change our conclusions if we did know it. Of course, sometimes we are wrong about this, and we acquire new knowledge that forces us to change our minds about some previously drawn conclusions. This is just to say that we reason *defeasibly*. Defeasible reasoning is reasoning that enables the reasoner to form beliefs provisionally, with the understanding that if more knowledge becomes available, the reasoner may have to retract some of his earlier conclusions.

How does defeasible reasoning work? It is a matter of reasoning, so it proceeds by drawing conclusions from previous conclusions by employing inference schemes. It differs from deductive reasoning in that the inference schemes employed do not guarantee the truth of the conclusion given the truth of the premises. The best the inference schemes can do is make the conclusion probable. We employ a wide array of defeasible inference schemes. For example, in order to gain knowledge of the world through perception, we must assume defeasibly that things tend to be the way they appear. In order to combine knowledge gained from perception at sometime different times, we must assume that the world tends to be stable — that what is true at one time tends to remain true at somewhat later times ("temporal projection"). In order to gain knowledge of probabilities, we must assume that observed relative frequencies tend to approximate actual probabilities ("statistical induction"). And so on.

It is tempting to think of all defeasible inferences as proceeding in accordance with the statistical syllogism:

> From "The probability is high of an arbitrary A being a B, and c is an A", infer defeasibly that c is a B.

This is an important defeasible inference scheme that we employ regularly. (Pollock 1983, 1990). However, we cannot reduce all defeasible reasoning to applications of this scheme, for the simple reason that agents do not start out knowing the values of the relevant probabilities. They must acquire knowledge of probabilities by gaining knowledge of individual facts by defeasible reasoning from perception, combine those facts by employing other defeasible inference schemes

like temporal projection, and then reason inductively from those individual facts to general probabilities. The upshot is that a cognitive agent must have a number of built-in defeasible inference schemes that he can employ before he acquires the knowledge of probabilities that is required for using the statistical syllogism.

Reasoning proceeds by stringing together defeasible (and deductive) reason schemes to produce arguments. Defeasible reasoning is more complex than deductive reasoning because, if an argument contains defeasible inferences, then a second argument might support defeaters for some of the inferences in the first argument. I assume the taxonomy of defeaters that I introduced in my (1970) and (1974) and that has been endorsed by most subsequent work on defeasible reasoning (see Prakken and Vreeswijk 2002 and Chesñevar, Maguitman, and Loui 2000). According to this taxonomy, there are two importantly different kinds of defeaters. Where P is a defeasible 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 there are other defeaters as well. For instance, suppose x looks red to me, but I know that x is illuminated by red lights and red lights can make objects look red when they are not. Knowing this defeats the defeasible reason, but it is not a reason for thinking that x is *not* red. After all, red objects look red in red light too. This is an *undercutting defeater*. Undercutting defeaters attack the *connection* between the reason and the conclusion rather than attacking the conclusion directly. For example, an undercutting defeater for the inference from x's looking red to x's being red attacks the connection between "x looks red to me" and "x is red", giving us a reason for doubting that x wouldn't look red unless it were red. I will symbolize the negation of "P wouldn't be true unless Q were true" as "$P \otimes Q$". A shorthand reading is "P does not guarantee Q". If Γ (a set of propositions) is a defeasible reason for P, then where $\Pi\Gamma$ is the conjunction of the members of Γ, any reason for believing "$\Pi\Gamma \otimes P$" is a defeater. Thus I propose to characterize undercutting defeaters as follows:

> If Γ is a defeasible reason for P, an *undercutting defeater* for Γ as a defeasible reason for P is any reason for believing "$(\Pi\Gamma \otimes P)$".

Are there any defeaters other than rebutting and undercutting defeaters? A number of authors have advocated what they call "specificity defeaters" (e.g., Touretzky 1984, Poole 1988, Simari and Loui 1992). However, I have argued at length (Pollock, 1995) that specificity defeat is unique to one kind of defeasible inference — that in accord with the statistical syllogism — and is not a general feature of defeasible reasoning, so it is not incorporated into OSCAR (although it is incorporated into the implementation of the statistical syllogism within OSCAR).

2. The Hard Problem

Given a set of arguments, some of which support defeaters for others, what should the reasoner believe? Unlike the case of deductive reasoning, he should not simply accept the conclusions of all of his arguments, because if some are accepted, others must be considered defeated. A theory of defeasible reasoning must tell us how to determine which conclusions to accept. An account of this is called a *semantics* for defeasible reasoning, although it need not be a semantics in the sense of model-theoretic semantics for formal logics. Constructing a satisfactory semantics has proven to be the hard problem for defeasible reasoning. This paper presents my latest thoughts on this matter.

We collect all of an agent's arguments into an *inference-graph*, where the nodes represent the conclusions of arguments, *support-links* tie nodes to the nodes from which they are inferred, and *defeat-links* indicate defeat relations between nodes. These links relate their *roots* to their *targets*. The root of a defeat-link is a single node, and the root of a support-link is a set of nodes. The analysis is somewhat simpler if we construct the inference-graph in such a way that when the same conclusion is supported by two or more arguments, it is represented by a separate node for each argument. For example, consider the inference-graph diagrammed in figure one, which represents two different arguments for $(P\&Q)$ given the premises, P, Q, A, and $(Q \rightarrow (P\&Q))$. The nodes of such an inference-graph represent arguments rather than just representing their conclusions. In such an inference-graph, a node has at most one support-link. When it is unambiguous to do so, I will refer to the nodes in terms of the conclusions they encode.

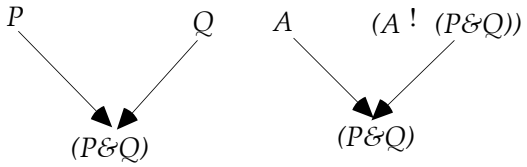


Figure 1. An inference-graph

The *node-basis* of a node is the set of roots of its support-links (if it has any), i.e., the set of nodes from which the node is inferred in a single step. If a node has no support-link (i.e., it is a premise) then the node-basis is empty. The *node-defeaters* are the roots of the defeat-links having the node as their target.

Given an inference-graph, a semantics must determine which nodes encode (the conclusions of) arguments that ought to be accepted, i.e., that are not defeated. This is the *defeat-status computation*, and nodes are marked "defeated" or "undefeated". The defeat-status computation is made more complex by the fact that some arguments support their conclusions more strongly than other arguments. For instance, if Jones tells me it is raining, and Smith denies it, and I regard them as equally reliable, then I have equally strong arguments both for believing that it is raining and for believing that it is not raining. In that case, I should withhold belief, not accepting either conclusion. On the other hand, if I regard Jones as much reliable than Smith, then I have a stronger argument for believing that it is raining, and if the difference is great enough, that is the conclusion I should draw. So argument-strengths make a difference. However, most semantics for defeasible reasoning ignore argument strengths, supposing that all premises are equally well justified and all inference schemes equally strong. One of the objectives of this chapter is to produce a semantics that takes account of strengths, but let us begin with this simplifying assumption that all arguments are equally strong. What can we say about the semantics in that simplified case?

Let us define:

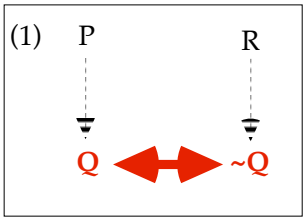
A node of the inference-graph is *initial* iff its node-basis and list of node-defeaters is empty.

It is initially tempting to try to characterize defeat statuses recursively using the following three rules:

(1) Initial nodes are undefeated.

(2) A non-initial node is undefeated if all the members of its node-basis are undefeated and all node-defeaters are defeated.

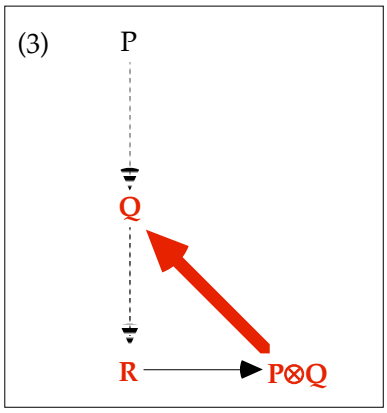
However, this recursion turns out to be ungrounded because we can have nodes of an inference-graph that defeat each other, as in inference-graph (1), where dashed arrows indicate defeasible inferences and heavy arrows indicate defeat-links. In computing defeat statuses in inference-graph (1), we cannot proceed recursively using rules (1)–(2), because that would require us to know the defeat status of Q before computing that of $\sim Q$, and also to know the defeat status of $\sim Q$ before computing that of Q. The general problem is that a node Q can have an inference/defeat-descendant that is a defeater of Q, where an inference/defeat-descendant of a node is any node that can be reached from the first node by following support-links and defeat-links. I will say that a node is *Q-dependent* iff it is an inference/defeat-descendant of a node Q. So the recursion is blocked in inference-graph (1) by there being Q-dependent defeaters of Q and $\sim Q$-dependent defeaters of $\sim Q$.



Inference-graph (1) is a case of "collective defeat". For example, let *P* be "Jones says that it is raining", *R* be "Smith says that it is not raining", and *Q* be "It is raining". Given *P* and *Q*, and supposing you regard Smith and Jones as equally reliable, what should you believe about the weather? It seems clear that you should withhold belief, accepting neither. In other words, both *Q* and ~*Q* should be defeated. This constitutes a counter-example to rule (2). So not only do rules (1)–(2) not provide a recursive characterization of defeat statuses — they are not even true. The failure of these rules to provide a recursive characterization of defeat statuses suggests that no such characterization is possible, and that in turn suggested to me (in my 1994, 1995) that rules (1)–(2) might be used to characterize defeat statuses in another way. Reiter's (1980) default logic proceeded in terms of multiple "extensions", and "skeptical default logic" characterizes a conclusion as following nonmonotonically from a set of premises and defeasible inference-schemes iff it is true in every extension. There are simple examples showing that this semantics is inadequate for the general defeasible reasoning of epistemic agents (see section two), but the idea of having multiple extensions suggested to me that rules (1)–(2) might be used to characterize multiple "status assignments". On this approach, a partial status assignment is an assignment of defeat statuses to the nodes of the inference-graph in accordance with (1)–(2):

> 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 any initial node;
> 2. σ assigns "undefeated" to a non-initial node α iff σ assigns "undefeated" to all the members of the node-basis of α and all node-defeaters of α are assigned "defeated".

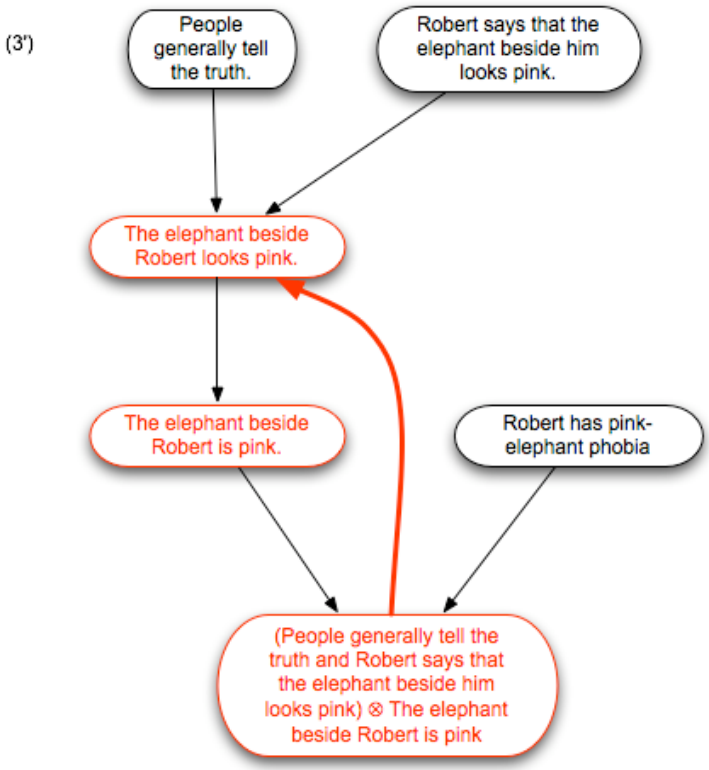
The reason for making status assignments "partial" is that there are inference graphs for which it is impossible to construct status assignments assigning statuses to every node. One case in which this happens is when we have "self-defeating arguments", i.e., arguments whose conclusions defeat some of the inferences leading to those conclusion. A simple example is inference-graph (3).



A partial status assignment must assign "undefeated" to *P*. If it assigned "undefeated" to *Q* then it would assign "undefeated" to *R* and *P*⊗*Q*, in which case it would have to assign "defeated" to *Q*. So it cannot assign "undefeated" to *Q*. If it assigned "defeated" to *Q* it would have to assign "defeated" to *R* and *P*⊗*Q*, in which case it would have to assign "undefeated" to *Q*. So that is not possible either. Thus a partial status assignment cannot assign anything to *Q*, *R*, and *P*⊗*Q*. Hence there is only one status assignment (i.e., maximal partial status assignment). Accordingly, *P* is undefeated and the other nodes are defeated. An intuitive example having approximately the same form is shown in inference-graph (3′). Here we suppose that people generally tell the truth, and this gives us a reason for believing what they tell us. However, some people suffer from a malady known as "pink-elephant phobia". In the presence of pink elephants, they become strangely disoriented so that their statements about their surroundings cease to be reliable. Now imagine Robert, who tells us that the elephant beside him looks pink. In ordinary circumstances, we would infer that the elephant beside Robert does look pink, and hence probably is pink. However, Robert suffers from pink-elephant phobia. So if it were true that the elephant beside Robert is pink, we could not rely

upon his report to conclude that it is. So we should not conclude that it is pink. We may be left wondering why he would say that it is, but we cannot explain his utterance by supposing that the elephant really is pink. So this gives us no reason at all for a judgment about the color of the elephant. On the other hand, it gives us no reason to doubt that Robert did say that the elephant is pink, or that Robert has pink-elephant phobia. Those are perfectly justified beliefs.

Inference-graphs (3) and (3′) constitute intuitive counterexamples to default logic (Reiter 1980) and the stable model semantics (Dung 1995) because there are no extensions. Hence on those semantics, *P* has the same status as *Q*, *R*, and *P*⊗*Q*. It is perhaps more obvious that this is a problem for those semantics if we imagine this self-defeating argument being embedded in a larger inference-graph containing a number of otherwise perfectly ordinary arguments. On these semantics, all of the nodes in all of the arguments would have to have the same status, because there would still be no extensions. But surely the presence of the self-defeating argument should not have the effect of defeating all other (unrelated) arguments.



To handle self-defeat, my (1995) semantics defined:

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

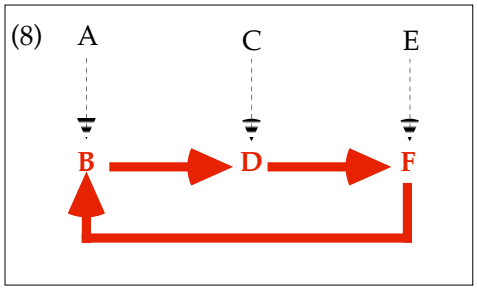
My proposal was then:

> A node is undefeated iff every status assignment assigns "undefeated" to it; otherwise it is defeated.
>
> Belief in P is justified for an agent iff P is encoded by an undefeated node of the inference-graph representing the agent's current epistemological state.

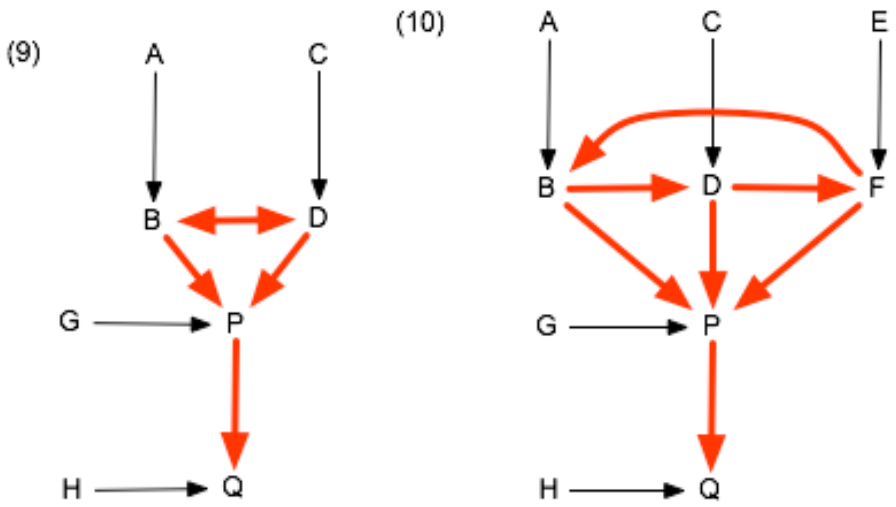
This has the consequence that in inference-graph (3), *P* is undefeated but *Q*, *R*, and *P*⊗*Q* are defeated. I take it that is a congenial result. I will refer to this semantics as *the multiple-assignment semantics*.

With the simplifying assumption that all arguments are equally strong, this is the semantics of

Pollock (1994, 1995). Both Prakken and Vreeswijk (2002) and Vo et al (2005) argue this semantics is equivalent to the subsequently developed preferred model semantics of Bonarenko et al (1997). This semantics produces the intuitively correct answer for many complicated inference-graphs. For example, consider inference-graph (1) again. The semantics produces two status assignments, one in which Q and $R \otimes S$ is assigned "defeated" and all other nodes are assigned "undefeated", and one in which $\sim Q$ is assigned "defeated" and all other nodes are assigned "undefeated". The result is that both Q and $\sim Q$ are defeated, and P and R are undefeated. For a number of years, I thought that, given the simplifying assumption, this semantics was correct. But I no longer think so. Here is the problem. Contrast inference-graph (1) with inference-graph (8). Inference-graph (8) involves "odd-length defeat cycles". For an example of inference-graph (8), let A = "Jones says that Smith is unreliable", B = "Smith is unreliable", C = "Smith says that Robinson is unreliable", D = "Robinson is unreliable", E = "Robinson says that Jones is unreliable", F = "Jones is unreliable". Intuitively, this should be another case of collective defeat, with A, C, and E being undefeated and B, D, and F being defeated. The multiple-assignment semantics does yield this result, but it does it in a peculiar way. A, C, and E must be assigned "undefeated", but there is no consistent way to assign defeat statuses of B, D, and F. Accordingly, there is only one status assignment (maximal partial status assignment), and it leaves B, D, and F unassigned. We get the right answer, but it seems puzzling that we get it in a different way than we do for even-length defeat cycles like that in inference-graph (1). This difference has always bothered me.



That we get the right answer in a different way does not show that the semantics is incorrect. As long as otherwise equivalent inference-graphs containing odd-length and even-length defeat cycles always produce the same defeat statuses throughout the graphs, there is no problem. However, they do not. Contrast inference-graphs (9) and (10). In inference-graph (9), there are two status-assignments, one assigning "defeated" to B and "undefeated" to D, and the other assigning "undefeated" to B and "defeated" to D. On either status assignment, P has an undefeated defeater, so it is defeated on both status assignments, with the result that Q is undefeated on both status-assignments. Hence Q is undefeated simpliciter. However, in inference-graph (10), there is only one status-assignment, and it assigns no status to any of B, D, F, P, or Q. Thus Q is defeated in inference-graph (10), but undefeated in inference-graph (9). This, I take it, is a problem. Although it might not be clear which inference-graph is producing the right answer, the right answer out to be the same for both inference-graphs. Thus one of them is getting it wrong. It is worth noting in passing that, as far as I know, no currently available semantics for defeasible reasoning handles (9) and (10) correctly. I take this to show that we need a different semantics.



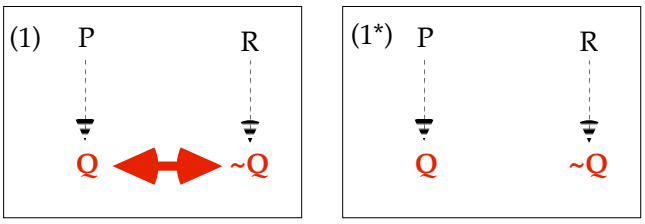
3. A Recursive Semantics

The multiple-assignment semantics is based upon the two rules:

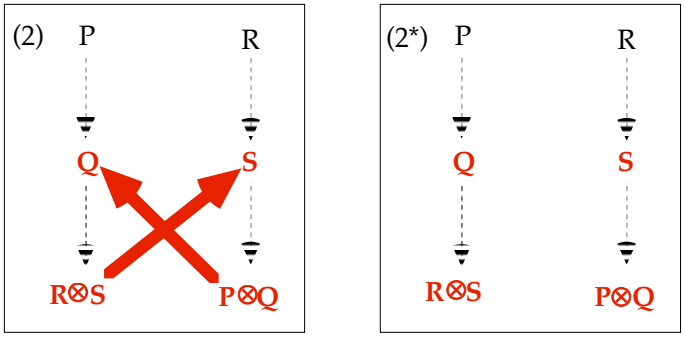
(1) Initial nodes are undefeated.

(2) A non-initial node is undefeated if all the members of its node-basis are undefeated and all node-defeaters are defeated.

We have seen that these rules are not true as stated. For example, inference-graph (1) is a counter-example to rule (2). Both *Q* and ~*Q* should be defeated, but then both have undefeated node-bases but no undefeated defeaters. I tried to avoid this problem by imposing these rules instead on partial-status assignments. But perhaps we should take seriously the fact that these rules are simply wrong. In inference-graph (2), in computing the defeat status of *Q*, what is crucial is that (a) its node-basis is undefeated, (b) the node-basis of its defeater is undefeated, and (c) there is no other defeater for ~*Q* besides *Q* itself. We can capture this by asking whether ~*Q* would be defeated if it were not defeated by *Q*. We can test this by removing the mutual defeat-links between *Q* and ~*Q*, producing inference-graph (1*). In (1*), ~*Q* is undefeated, and because of that, *Q* is defeated in (1). Note that the defeaters we are removing in constructing inference-graph (1*) are those that are *Q*-dependent, i.e., those that can be reached by following paths from *Q* consisting of inference-links and defeat-links.



Consider another example — inference-graph (2). In computing the defeat-status of *Q*, we note that its node-basis is undefeated, and its defeater *P*⊗*Q* is defeated only by the *Q*-dependent defeat-link from *R*⊗*S*. If we remove the *Q*-dependent defeat-links from inference-graph (2) we get inference-graph (2*). In inference-graph (2*), *P*⊗*Q* is undefeated, so this makes *Q* defeated in inference-graph (2).



These examples suggest that we might replace rule (2) by a rule that computes the defeat-statuses of defeat-links in a modified inference-graph from which we have removed those defeat-links that make the computation circular. Recall that a defeat-link or support-link extends from its *root* to its *target*. Let us define:

Definition: An *inference/defeat-path* from a node φ to a node θ is a sequence of support-links and defeat-links such that (1) φ is the root of the first link in the path; (2) θ is the last link in the path; (3) the root of each link after the first member of the path is the target of the preceding link; (4) the path does not contain an internal loop, i.e., no two links in the path have the same target.

Definition: θ is φ-*dependent* iff there is an inference/defeat-path from φ to θ.

Definition: A *circular inference/defeat-path* from a node φ to itself is an inference/defeat-path from φ to a defeat-link for φ.

Definition: A defeat-link is φ-critical iff it is a member of some minimal set of defeat-links such that removing all the defeat-links in the set suffices to cut all the circular inference/defeat-paths from φ to φ.

It will be convenient to modify our understanding of initial nodes. Previously, I took them to be automatically undefeated, and we can still regard that as the default value, but it will also be useful to be able to stipulate that some of the initial nodes in a newly-constructed inference-graph are defeated. This allows us to define:

Definition: If φ is a node of an inference-graph *G*, let G_φ be the inference-graph that results from deleting all φ-critical defeat-links from *G* and making all members of the node-basis of φ and all φ-independent nodes initial-nodes (i.e., delete their support-links and defeat-links) with stipulated defeat-statuses the same as their defeat-statuses in *G*.

My proposed semantics now consists of two rules:

(DS1) Initial nodes are undefeated unless they are stipulated to be defeated.

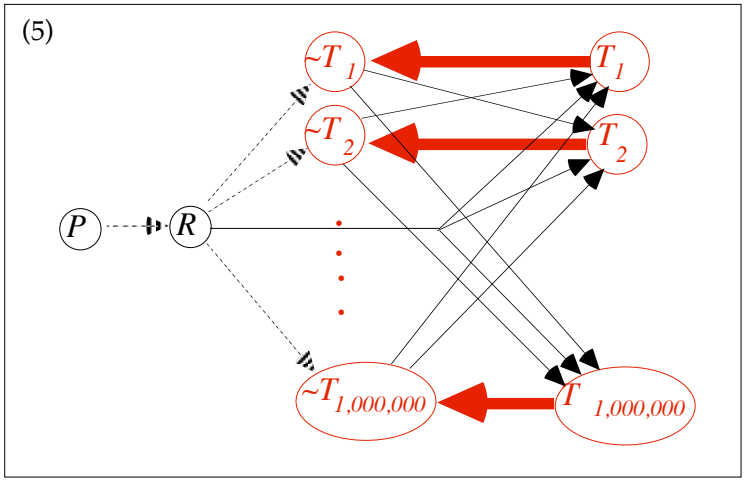
(DS2) A non-initial node φ is undefeated in an inference-graph *G* iff all members of the node-basis of φ are undefeated in *G* and any defeater for φ is defeated in G_φ.

On the assumption that arguments cannot be circular, this pair of rules can be applied recursively to compute the defeat-status of any node in a finite inference-graph. The recursion simply steps through arguments, computing the defeat-status of each node φ after the defeat-statuses of the nodes in φ's node-basis are computed. The problem of circular inference/defeat-paths is avoided by

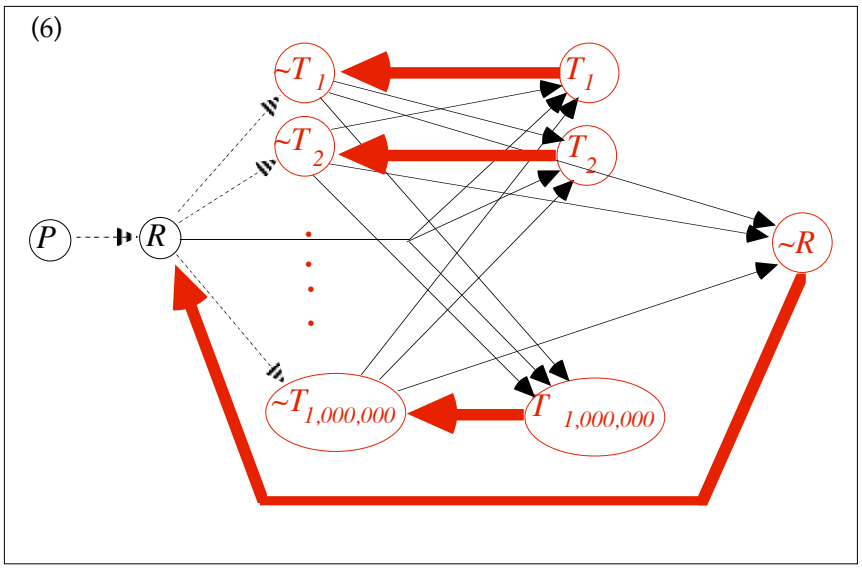
simply removing the φ-critical defeat-links and evaluating node-defeaters in G_φ. I will refer to this new semantics as *the recursive semantics*, and constrast it with the multiple-status-assignment semantics.

I believe that the recursive semantics gets everything right that the multiple-assignment semantics got right. Consider a more complex example. Inference-graph (5) illustrates the so-called "lottery paradox" (Kyburg 1961). Here P reports a description (e.g., a newspaper report) of a fair lottery with one million tickets. P constitutes a defeasible reason for R, which is the description. In such a lottery, each ticket has a probability of one in a million of being drawn, so for each i, the statistical syllogism gives us a reason for believing $\sim T_i$ ("ticket i will not be drawn"). The supposed paradox is that although we thusly have a reason for believing of each ticket that it will not be drawn, we can also infer on the basis of R that some ticket will be drawn. Of course, this is not really a paradox, because the inferences are defeasible and this is a case of collective defeat. This results from the fact that for each i, we can infer T_i from the description R (which entails that some ticket will be drawn) and the conclusions that none of the other tickets will be drawn. This gives us a defeating argument for the defeasible argument to the conclusion that $\sim T_i$, as diagrammed in inference-graph (5). The result is that for each i, there is a status assignment on which $\sim T_i$ is assigned "defeated" and the other $\sim T_j$'s are all assigned "undefeated", and hence none of them are assigned "undefeated" in every status assignment.

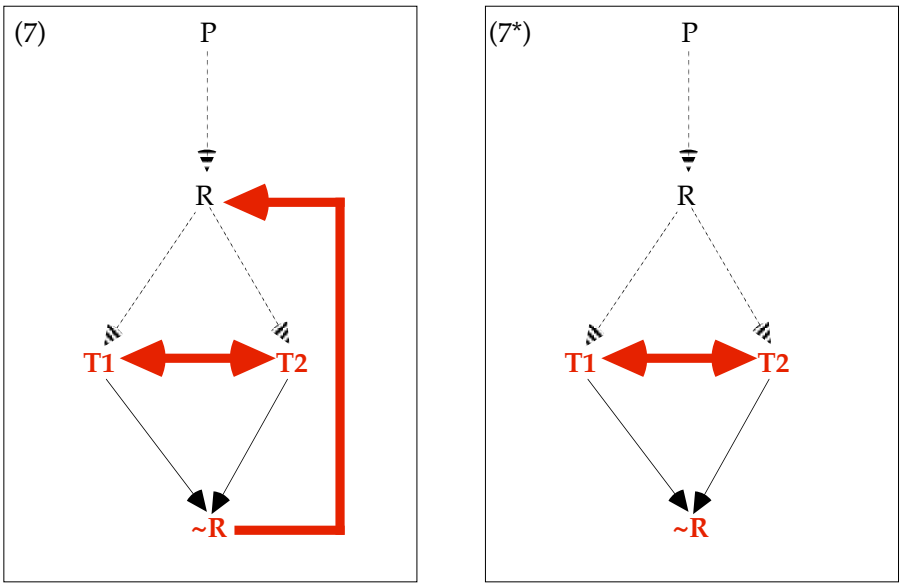
(5)



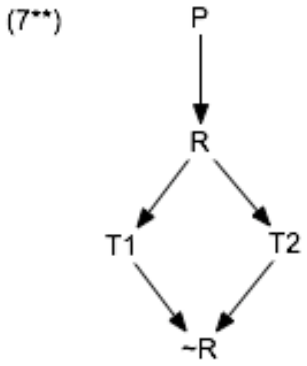
I believe that all (skeptical) semantics for defeasible reasoning get the lottery paradox right. A more interesting example is the "lottery paradox *paradox*", diagrammed in inference-graph (6). This results from the observation that because R entails that some ticket will be drawn, from the collection of conclusions of the form $\sim T_i$ we can infer $\sim R$, and that is a defeater for the defeasible inference from P to R. This is another kind of self-defeating argument. Clearly, the inferences in the lottery paradox should not lead us to disbelieve the newspaper's description of the lottery, so R should be undefeated. Circumscription (McCarthy 1986), in its simple non-prioritized form, gets this example wrong, because one way of minimizing abnormalities would be to block the inference from P to R. My own early analysis (Pollock 1987) also gets this wrong. This was the example that led me to the analysis of section one. That analysis gets this right. We still have the same status assignments as in inference-graph (5), and $\sim R$ is defeated in all of them because it is inferred from the entire set of $\sim T_i$'s, and one of those is defeated in every status assignment.



It will be convenient to have a simpler example of an inference-graph with the same general structure as the lottery paradox paradox. For that purpose we can use inference-graph (7). Here *P* and *R* should be undefeated, but *T1*, *T2*, and ~*R* should be defeated. To compute the defeat-status of *R* in inference-graph (7), we construct (7*) by removing the only defeat-link whose removal results in *R* no longer having an *R*-dependent defeater. In (7*), the triangle consisting of *R*, *T1* and *T2* is analogous to inference-graph (1), with the result that T_1 and *T2* are both defeated in inference-graph (7*). They constitute the node-basis for ~*R*, so ~*R* is also defeated in inference-graph (7*). Thus *R* is undefeated in inference-graph (7). Turning to T_1 and T2 in inference-graph (7), both have *R* as their node-basis, and *R* is defeated, so they are defeated. Then because T_1 and T2 are defeated, ~*R* is defeated in inference-graph (7). So we get the intuitively correct answers throughtout.

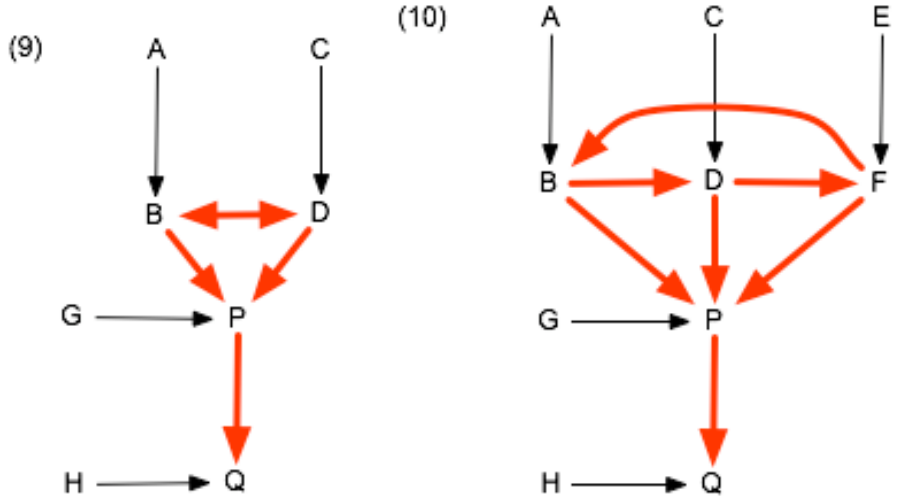


Inference-graph (7) also illustrates why, in constructing G_{φ}, we remove only the φ-critical defeat-links, and not all of the φ-dependent defeat-links. All of the defeat-links in inference-graph (7) are *R*-dependent, and if we remove them all we get inference-graph (7**). But in inference-graph (7**), ~R is undefeated. This would result in *R* being defeated in inference-graph (7) rather than undefeated. Thus it is crucial to remove only the φ-critical defeat-links rather than all the φ-dependent defeat-links.

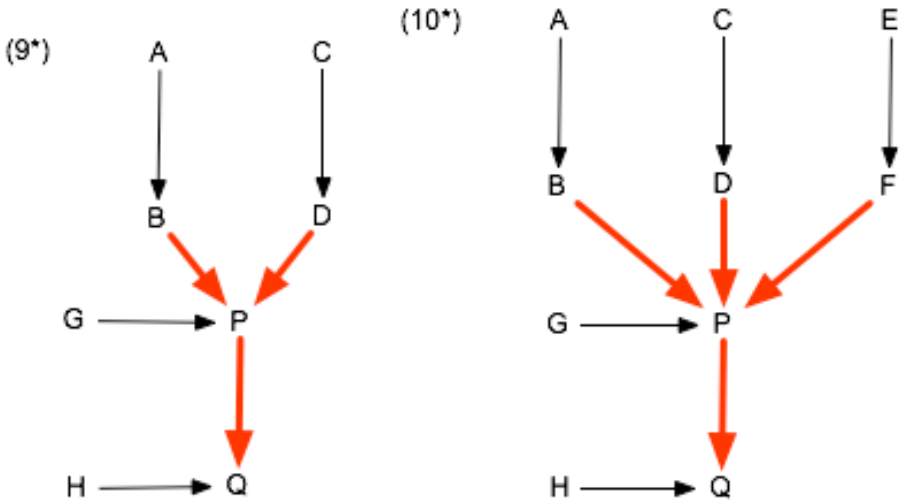


4. The Problem Cases

Now let us turn to some cases that the multiple-assignment semantics does not or may not get right. First, consider the pair of inference-graphs that motivated the search for a new semantics. These are inference-graphs (9) and (10). In these inference-graphs, not everyone agrees whether *Q* should come out defeated or undefeated, but it does seem clear that whatever the right answer is, it should be the same for both inference-graphs. Unfortunately, on the multiple-assignment semantics, *Q* is undefeated in inference-graph (9) and defeated in inference-graph (10).



On the (new) recursive semantics, we compute the defeat-statuses of *B* and *D* in inference-graph (9) by constructing inference-graph (9*). *B* and *D* are undefeated in inference-graph (9*), so each defeats the other in inference-graph (9), with the result that *B* and *D* are defeated in inference-graph (9). There are no *P*-critical defeat-links in (9), so removing *P*-critical defeat-links leaves inference-graph (9) unchanged. *B* and *D* are defeated in inference-graph (9), so it follows that *P* is defeated in inference-graph (9). Then because there are no *Q*-dependent defeat-links in inference-graph (9), *Q* is undefeated.



The computation of defeat-statuses in inference-graph (10) works in exactly the same way, via inference-graph (10*), again producing the result that *Q* is undefeated. So on the recursive semantics, we do not get a divergence between inference-graphs (9) and (10).

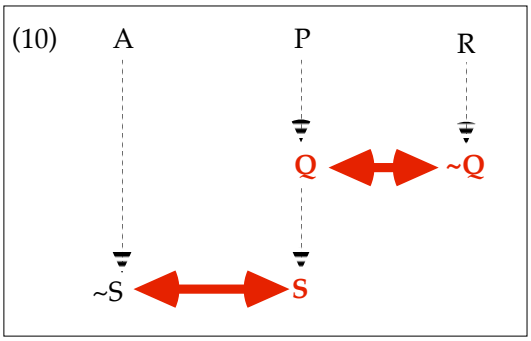
Still, we can ask whether the answer we get for inference-graphs (9) and (10) is the correct answer. There is some intuitive reason for thinking so. In inference-graph (9), *B* and *D* are defeated, so they should not have the power to defeat *P*, and hence *P* should defeat *Q*. Similarly, in inference-graph (10), all three of *B*, *D*, and *F* are defeated, and so again, *D* should not have the power to defeat *P*, and hence *P* should defeat *Q*.

Whether this is right is closely connected with a question that has puzzled theorists since the earliest work on the semantics of defeasible reasoning. The multiple-assignment semantics, as well as default logic, the stable model semantics, circumscription, and almost all standard semantics for defeasible reasoning and nonmonotonic logic, support what I have called (1987) "presumptive defeat".[1] For example, consider inference-graph (10). On the multiple-assignment semantics, a defeated conclusion like *Q* that is assigned "defeated" in some status assignment and "undefeated" in another retains the ability to defeat. That is because, in the assignment in which it is undefeated, the defeatee is defeated, and hence not undefeated in all status-assignments. In the case of inference-graph (10) this has the consequence that *S* is assigned "defeated" in those status-assignments in which *Q* is assigned "defeated", but *S* is assigned "undefeated" and ~*S* is assigned "defeated" in those status-assignments in which *Q* is assigned "undefeated". Touretzky, Horty, and Thomason (1987) called this "ambiguity propagation", and Makinson and Schlechta (1991) called such arguments "Zombie arguments" (they are dead, but they can still get you). However, the recursive semantics precludes presumptive defeat. It entails *Q*, ~*Q*, and hence *S*, are all defeated, and ~*S* is undefeated. Is this the right answer? Consider an example. You are sitting with Keith and Alvin, and the following conversation ensues:

> Keith: I heard on the news this morning that it is going to rain this afternoon.
>
> Alvin: Nonsense! I was sitting right beside you listening to the same weather report, and the announcer clearly said that it is going to be a sunny day in Tucson.
>
> Keith: You idiot, you must have cotton in your ears! It was perfectly clear that he said it is going ro rain.
>
> Alvin: You never pay attention. No one in his right mind could have thought he said it was going to rain. He said it would be sunny.
>
> …

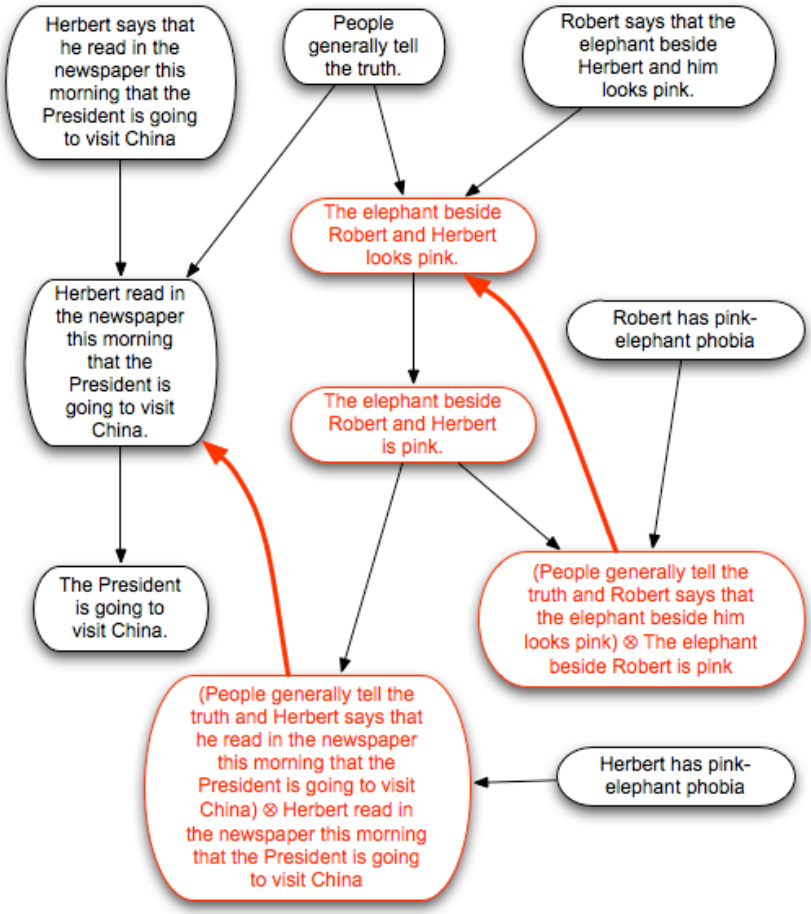
[1] The only semantics I know about that does not support presumptive defeat are certain versions of Nute's (1992) defeasible logic. See also Covington, Nute, and Vellino (1997), and Nute (1999).

At that point, you wander off shaking your head, still wondering what the weather is going to be. Then it occurs to you that it is about time for the noon News, so you turn on the radio and hear the announcer say, "This just in from the National Weather Service. It is going to rain in Tucson this afternoon." Surely, that settles the matter. You will believe, with complete justification, that it is going to rain. The earlier conversation between Keith and Alvin does not defeat your judgment on the basis of the noon broadcast.



Presumptive defeat arises from the fact that if a node *P* is defeated in one assignment and undefeated in another, then all *P-dependent* nodes will also have different defeat statuses in the different assignments unless one of their inference-ancestors is defeated absolutely (i.e., in all status assignments). A similar problem arises for inference-nodes *P* that cannot be assigned defeat-statuses in any assignments. This occurs, for example, in cases of self-defeat or when there are odd-length defeat cycles. In this case, no *P*-dependent node can be assigned defeat statuses either unless one of their inference-ancestors is defeated absolutely. For example, consider once more the sad case of Robert, the pink-elephant-phobic. We observed that Robert's statement that the elephant beside him is pink does not give us a good reason for believing that it really is pink. Now suppose that Robert is accompanied by Herbert, who is also standing beside the elephant. While Robert is blathering about pink-elephants, Herbert turns to you and say, "I read in the newspaper this morning that the President is going to visit China." From this you infer that he did read that in the newspaper, and hence the President is probably going to visit China. Suppose, however, that Herbert also suffers from pink-elephant-phobia. Does that make any difference? It does not seem so, because as we observed, Robert's statement gives us no reason to think the elephant is pink, and so no reason to distrust Herbert's statement. This scenario is diagrammed in figure 12. However, on the multiple-assignment semantics, "The elephant beside Robert and Herbert is pink" has no status assignment, and hence neither does "(People generally tell the truth and Herbert says that he read in the newspaper this morning that the President is going to visit China) ⊗ Herbert read in the newspaper this morning that the President is going to visit China" or "Herbert read in the newspaper this morning that the President is going to visit China" or "The president is going to visit China". This seems clearly wrong. On the other hand, on the recursive semantics, "The elephant beside Robert and Herbert looks link" is defeated, and hence so is "The elephant beside Robert and Herbert is pink" and so is "(People generally tell the truth and Herbert says that he read in the newspaper this morning that the President is going to visit China) ⊗ Herbert read in the newspaper this morning that the President is going to visit China". Accordingly, "Herbert read in the newspaper this morning that the President is going to visit China" and "The president is going to visit China" are undefeated, which is the intuitively correct result.

(12)



The upshot is that the recursive semantics agrees with the older multiple-assignment semantics on simple cases in which the latter seems to give the right answer, but the recursive semantics also seems to get right a number of cases that the multiple-assignment semantics gets wrong. The test of a semantics for defeasible reasoning is that it agrees with our intuitions about clear cases. So we have reasonably strong inductive reasons for thinking the recursive semantics properly characterizes correct inference in defeasible reasoning.

5. Computing Defeat-Statuses

Principles (DS1) and (DS2) provide a recursive characterization of defeat-status relative to an inference-graph. However, this characterization does not lend itself well to implementation because it requires the construction of modified inference-graphs, which would be computationally expensive. The objective of this section is to produce an equivalent recursive characterization that appeals only to the given inference-graph.

A defeat-link is φ-critical iff it is a member of a minimal set such that removing all the defeat-links in the set suffices to cut all the circular inference/defeat-paths from φ to φ. A necessary condition for a defeat-link L to be φ–critical is that it lie on such a circular path. In general, there can be diverging and reconverging paths with several "parallel" defeat-links, as in figure 6. In figure 6, removing the defeat-link D_1 suffices to cut both circular paths. But the set $\{D_1,D_2\}$ of parallel defeat-links is also a minimal set of defeat-links such that the removal of all the links in the set suffices to cut all the circular inference/defeat-paths from φ to φ. Thus in figure 6, all of the defeat-links are φ–critical. However, lying on a circular inference/defeat-path is not a sufficient condition for being φ-critical. A defeat-link on a circular inference/defeat-path from φ to φ can fail to be φ–critical is when there is a path around it consisting entirely of support-links, as diagrammed in figure 7. In this case, you must remove D_1 to cut both paths, and once you have done that, removing D_2 is a

gratuitous additional deletion. So D_2 is not contained in a minimal set of deletions sufficient for cutting all the circular inference/defeat-paths from φ to φ, and hence D_2 is not φ-critical. This phenomenon is also illustrated by inference-graph (7), and it is crucial to the computation of degrees of justification in that inference-graph that such defeat-links not be regarded as φ-critical. It turns out that this is the only way a defeat-link on a circular inference/defeat-path can fail to be φ-critical, as will now be proven.

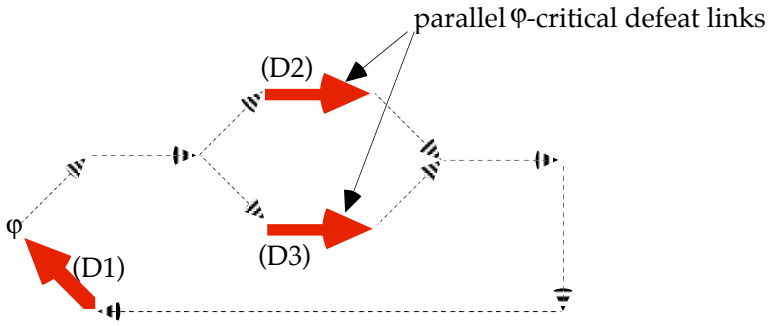


Figure 6. Parallel φ-critical defeat-links

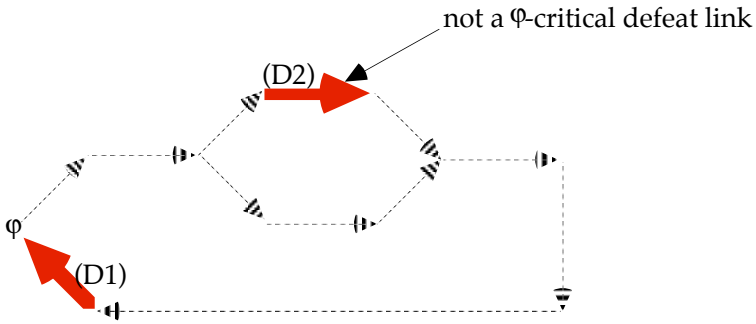


Figure 7. Defeat link that is not φ-critical

Let us say that a node α *precedes* a node β on an inference/defeat-path iff α and β both lie on the path and either α = β or the path contains a subpath originating on α and terminating on β. *Node-ancestors* of a node are nodes that can be reached by following support-links backwards. It will be convenient to define:

> **Definition:** A defeat-link *L can be bypassed* on an inference/defeat-path μ in *G* iff there is a node α preceding the root of *L* on μ and a node β preceded by the target of *L* on μ such that α = β or α is a node-ancestor of β in *G*.

It will be convenient to define:

> **Definition:** μ is a φ-*circular-path* in *G* iff μ is a circular inference/defeat-path in *G* from φ to φ and no defeat-link in *G* can be bypassed.

I will prove the following theorem, which is of central importance in implementing the theory of defeasible reasoning.

Theorem 40: A defeat-link is φ-critical in *G* iff it lies on a φ-circular-path in *G*.

This theorem follows immediately from the next three lemmas.

Lemma 41: If μ_1 and μ_2 are φ-circular-paths and every defeat-link in μ_1 occurs in μ_2, then μ_1 and μ_2 contain the same defeat-links and they occur in the same order.

Proof: Suppose the defeat-links in μ_1 are $\delta_1,\ldots,\delta_n$, occurring in that order. Suppose μ_1 and μ_2 differ first at the ith defeat-link. Then μ_1 and μ_2 look as in figure 8. But every defeat-link in μ_1 occurs in μ_2, so δ_i must occur later in μ_2. But then there is a bypass around δ_i* in μ_2, which is impossible if it is a φ-circular-path. ■

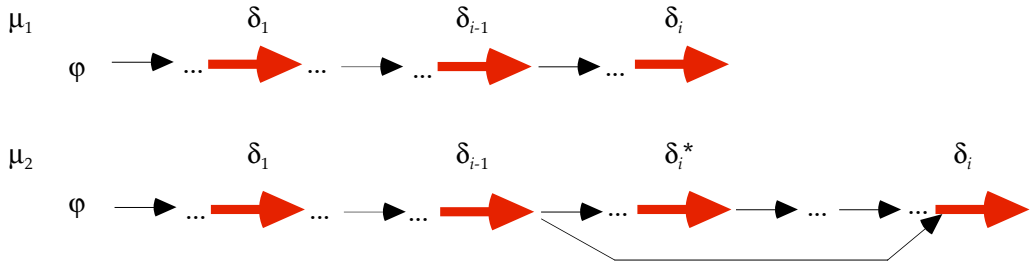


Figure 8. Paths must agree.

Lemma 42: Every defeat-link in a φ-circular-path is φ-critical.

Proof: Suppose δ is a defeat-link on the φ-circular-path μ. Let D be the set of all defeaters other than those on μ. If deleting all members of D is sufficient to cut all φ-circular-paths not containing δ, then select a minimal subset D_0 of D whose deletion is sufficient to cut all φ-circular-paths not containing δ. Adding δ to D_0 gives us a set of defeat-links whose deletion is sufficient to cut all φ-circular-paths. Furtheremore, it is minimal, because adding δ cannot cut any paths not containing δ, and all members of D_0 are required to cut those paths. Thus δ is a member of a minimal set of defeat-links the deletion of which is sufficient to to cut all φ-circular-paths, i.e., δ is φ-critical.

Thus if δ is not φ-critical, there is a φ-circular-path ν not containing δ and not cut by cutting all defeat-links not in μ. That is only possible if every defeat-link in ν is in μ. But then by the previous lemma, μ and ν must contain the same defeat-links, so contrary to supposition δ is in ν. So the supposition that δ is not φ-critical is inconsistent with the supposition that it lies on a φ-circular-path. ■

Lemma 43: If a defeat-link does not occur on any φ-circular-path then it is not φ-critical.

Proof: For every circular inference/defeat-path μ from φ to φ there is a φ-circular-path ν such that every defeat-link in ν is in μ. ν results from removing bypassed defeat-links and support-links in μ and replacing them by their bypasses. It follows that any set of deletions of defeat-links that will cut all φ-circular-paths will also cut every circular inference/defeat-path from φ to φ. Conversely, φ-circular-paths are also circular-paths from φ to φ, so any set of deletions that cuts all circular-paths from φ to φ will also cut all φ-circular-paths. So the φ-circular-paths and the circular-paths from φ to φ have the same sets of deletions of defeat-links sufficient to cut them, and hence the same minimal sets of deletions. If a defeat-link δ does not occur on any φ-circular-path, then it is irrelevant to cutting all the φ-circular-paths, and hence it is not in any minimal set of deletions sufficient to cut all circular-paths from φ to φ, i.e., it is not φ-critical. ■

Theorem 40 follows immediately from lemmas 42 and 43.

A further simplification results from observing that, for the purpose of deciding whether a defeat-link is φ-critical, all we have to know about φ-circular-paths is what defeat-links occur in them. It makes no difference what support-links they contain. So let us define:

Definition: A φ-*defeat-loop* is a sequence μ of defeat-links for which there is a φ-circular-path ν such that the same defeat-links occur in μ and ν and in the same order.

In other words, to construct a φ-defeat-loop from a φ-circular-path we simply remove all the support-links. We have the following very simple characterization of φ-defeat-loops:

Theorem 44: A sequence $\langle\delta_1,...,\delta_n\rangle$ of defeat-links is a φ-defeat-loop iff (1) φ is a node-ancestor of the root of δ_1 but not of the root of any δ_k for $k > 1$, (2) φ is the target of δ_n, and (3) for each $k < n$, the target of δ_k is equal to or an ancestor of the root of δ_{k+1}, but not of the root of δ_{k+j} for $j > 1$.

The significance of φ-defeat-loops is that by omitting the support-links we make them easier to process, but we still have the simple theorem:

Theorem 45: A defeat-link is φ-critical in *G* iff it lies on a φ-defeat-loop in *G*.

In simple cases, G_φ will be an inference-graph in which no node ψ has a ψ-critical defeat-link. But in more complex cases, like inference-graph (7), we have to repeat the construction, constructing first G_φ, and then $(G_\varphi)_\psi$. Let us define recursively:

Definition: $G_{\langle\varphi_1,...,\varphi_n\rangle} = \left(G_{\langle\varphi_2,...,\varphi_n\rangle}\right)_{\varphi_1}$

As formulated, the recursive semantics requires us to construct the inference-graphs $G_{\langle\varphi_1,...,\varphi_n\rangle}$. To reformulate the semantics so as to avoid this, let us define recursively:

Definition:

A defeat-link δ of *G* is $\langle\varphi_1,...,\varphi_n\rangle$-*critical* in *G* iff (1) δ lies on a φ_1-defeat-loop μ in *G* containing no $\langle\varphi_2,...,\varphi_n\rangle$-critical defeat-links.

A defeat-link δ of *G* is *hereditarily*-$\langle\varphi_1,...,\varphi_n\rangle$-*critical* in *G* iff either δ is $\langle\varphi_1,...,\varphi_n\rangle$-*critical* in *G* or δ is hereditarily-$\langle\varphi_2,...,\varphi_n\rangle$-*critical* in *G*.

A defeater (i.e., a node) of *G* is *hereditarily*-$\langle\varphi_1,...,\varphi_n\rangle$-*critical* in *G* iff it is the root of a hereditarily-$\langle\varphi_1,...,\varphi_n\rangle$-critical defeat-link in *G*.

Obviously:

Theorem 46: δ is hereditarily-$\langle\varphi_1,...,\varphi_n\rangle$-critical in *G* iff δ is φ_1-critical in $G_{\langle\varphi_2,...,\varphi_n\rangle}$ or φ_2-critical in

$G_{\langle \varphi_3,\ldots,\varphi_n \rangle}$ or ... or φ_n-critical in G.

Note that a defeat-link that is φ_i-critical in $G_{\langle \varphi_{i+1},\ldots,\varphi_n \rangle}$ does not exist in $G_{\langle \varphi_{j+1},\ldots,\varphi_n \rangle}$ for $j < i$, so:

Theorem 47: δ is φ_1-critical in $G_{\langle \varphi_2,\ldots,\varphi_n \rangle}$ iff δ is $\langle \varphi_1,\ldots,\varphi_n \rangle$-critical in G.

Furthermore, a defeat-link still exists in $G_{\langle \varphi_3,\ldots,\varphi_n \rangle}$ (i.e., has not been removed) iff it is not $\langle \varphi_1,\ldots,\varphi_n \rangle$-critical in G.

Where $\theta,\varphi_1,\ldots,\varphi_n$ are nodes of an inference-graph G, define:

Definition:

θ is $\langle \varphi \rangle$-*independent* of ψ in G iff there is no inference/defeat-path in G from φ to θ.

θ is $\langle \varphi_1,\ldots,\varphi_n \rangle$-*independent* in G iff every inference/defeat-path in G from φ_1 to θ contains a hereditarily-$\langle \varphi_2,\ldots,\varphi_n \rangle$-critical defeat-link.

Theorem 48: θ is $\langle \varphi_1,\ldots,\varphi_n \rangle$-*independent* in G iff θ is φ_1-independent in $G_{\langle \varphi_2,\ldots,\varphi_n \rangle}$.

Where ψ is an initial node in G, let $j_0(\psi,G)$ be its assigned value. Let us define recursively:

Definition:

(a) If ψ is initial in G then ψ is $\langle \varphi_1,\ldots,\varphi_n \rangle$-undefeated in G iff ψ is undefeated in G;

(b) If ψ is $\langle \varphi_1,\ldots,\varphi_n \rangle$-*independent* in G then ψ is $\langle \varphi_1,\ldots,\varphi_n \rangle$-undefeated G iff ψ is $\langle \varphi_2,\ldots,\varphi_n \rangle$-undefeated G;

(c) Otherwise, ψ is $\langle \varphi_1,\ldots,\varphi_n \rangle$-undefeated in G iff (1) all members of the node-basis of ψ are $\langle \varphi_1,\ldots,\varphi_n \rangle$-undefeated in G, (2) all defeaters for ψ that are $\langle \varphi_1,\ldots,\varphi_n \rangle$-independent of ψ in G and are not hereditarily-$\langle \varphi_1,\ldots,\varphi_n \rangle$-critical in G (i.e., still exist in $G_{\langle \varphi_3,\ldots,\varphi_n \rangle}$) are $\langle \varphi_1,\ldots,\varphi_n \rangle$-defeated in G, and (3) all defeaters for ψ that are $\langle \varphi_1,\ldots,\varphi_n \rangle$-dependent of ψ in G and are not hereditarily-$\langle \varphi_1,\ldots,\varphi_n \rangle$-critical in G (i.e., still exist in $G_{\langle \varphi_3,\ldots,\varphi_n \rangle}$) are $\langle \psi,\varphi_1,\ldots,\varphi_n \rangle$-defeated in G,

The reason this is a recursive definition is that we always reach an n at which there are no more $\langle \varphi_1,\ldots,\varphi_n \rangle$-dependent defeaters, and then the values of all nodes are computed recursively in terms of the values assigned to initial nodes.

It is now trivial to prove by induction on n that:

Theorem 49: ψ is undefeated in $G_{\langle \varphi_3,\ldots,\varphi_n \rangle}$ iff ψ is $\langle \varphi_1,\ldots,\varphi_n \rangle$-undefeated in G.

Thus we have a recursive definition of the degree of justification of a node that computes the degrees of justification entirely by reference to the given inference-graph rather than by building a sequence of modified inference-graphs in accordance with the original analysis.