Degrees of Justification

John L. Pollock
Department of Philosophy
University of Arizona
Tucson, Arizona 85721
(e-mail: pollock@arizona.edu)

Abstract

The question addressed in this paper is how the degree of justification of a belief is determined. This is made more complicated by the fact that a conclusion may be supported by several different arguments, the arguments typically being defeasible, and there may also be arguments of varying strengths for defeaters for some of the supporting arguments. What is sought is a way of computing the "on sum" degree of justification of a conclusion in terms of the degrees of justification of all relevant premises and the strengths of all relevant reasons.

I have in the past defended various principles pertaining to this problem. In this paper I reaffirm some of those principles but propose a significantly different final calculation. Specifically, I endorse the weakest link principle for the computation of argument strengths. According to this principle the degree of justification an argument confers on its conclusion in the absence of other relevant arguments is the minimum of the degrees of justification of its premises and the strengths of the reasons employed in the argument. I reaffirm my earlier rejection of the accrual of reasons, according to which two arguments for a conclusion can result in a higher degree of justification than either argument by itself. In the past I have also denied the accrual of defeat, according to which multiple defeaters can result in the defeat of an argument that is not defeated by any of the defeaters individually. In this paper I urge that there are compelling examples that support a limited version of the accrual of defeat. Although multiple rebutting defeaters or multiple undercutting defeaters provide no more defeat than the individual defeaters, a combination of rebutting and undercutting defeaters can defeat an argument that is not defeated by the individual defeaters.

The accrual of defeat has important consequences for the computation of degrees of justification. The paper closes with a proposal for computing degrees of justification that captures the various principles endorsed and seems to provide intuitively correct degrees of justification in the complex cases that motivated those principles.

Keywords: defeasible, defeat, justification, reasoning

This work was supported by NSF grant no. IRI-9634106.

An early version of this paper was presented to the 20th International Wittgenstein Symposium, and published as my (1998a). The conclusions of the present paper are importantly different, however.

1. Introduction

I have argued at length elsewhere that a rational agent operating in a complex environment must reason about its environment defeasibly.[1] The OSCAR architecture for rational agents is based upon this contention, and implements the system of defeasible reasoning described in Pollock (1995). The basic idea is that the agent constructs arguments using both deductive and defeasible reason-schemas (inference-schemas). The conclusions of some of these arguments may constitute defeaters for steps of some of the other arguments. Given the set of interacting arguments that represent the agent's epistemological state at a given time, an algorithm is run to compute defeat-statuses, determining which arguments are defeated and which are undefeated. What the agent should believe at that time are the conclusions of the undefeated arguments.

The literature on defeasible and nonmonotonic reasoning contains numerous proposals for how to compute defeat-statuses. OSCAR computes them in the manner described in my (1994) and (1995).[2] If we ignore the fact that some arguments provide stronger support for their conclusions than other arguments, we can describe OSCAR's defeat-status computation as follows. We collect 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. 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. So the nodes represent arguments rather than just representing their conclusions. The *node-basis* of a node is the set of nodes from which the node is inferred in a single step. We define:

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

A status assignment is then an assignment of defeat-statuses to the nodes of the inference graph in accordance with three simple rules:

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

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

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 supported by an undefeated argument of the inference-graph representing the agent's current epistemological state.

---

[1] The argument spans three decades. My most recent paper in this vein is Pollock (1998), but see also Pollock (1974), (1987), (1990), (1995), and (1999).

[2] For comparison with default logic and circumscription, see my (1995), chapter three. For comparison with more recent systems of defeasible argumentation, see Praaken and Vreeswijk (forthcoming).

This account of defeat-status assumes that all arguments support their conclusions equally strongly. However, this assumption is unrealistic. For example, increasing the degrees of justification of the premises of an argument may increase the degree of justification of the conclusion, and increasing the strengths of the reasons employed in the argument may increase the degree of justification of the conclusion. This phenomenon has been ignored in most AI work on defeasible and nonmonotonic reasoning, but it is of some importance in applications of such reasoning. For example, in Pollock (1998) I discussed temporal projection, wherein it is assumed defeasibly that if something is true at one time it is still true at a later time. This is, in effect, a default assumption that fluents are stable, tending not to change truth values as time passes. The stability of a fluent is measured by the probability ρ that if it is true at time t then it is still true at time t+1. More generally, if it is true at time t, then the probability of its being true at $t+\Delta t$ is $\rho^{\Delta t}$. So the strength of the defeasible expectation supported by temporal projection is a monotonic decreasing function of the time interval. This can be captured in a system of defeasible reasoning by employing a reasoning schema of the following sort:

> "P-at-t" is a defeasible reason for believing "P-at-$(t+\Delta t)$", the strength of the reason being a monotonic decreasing function of Δt.[3]

The decreasing strength is important in understanding perceptual updating, wherein on the basis of new perception the agent overrides temporal projection and concludes that the fluent has changed truth value. Perception is not infallible, so perception should provide only a defeasible reason for believing that the environment is as represented by the percept.[4] So suppose an object looks red at one time t_1 and blue at a later time t_2. The agent should assume defeasibly that the object is initially red, but should also conclude defeasibly that it changes color later and is blue at t_2. The object's being red at t_1 provides a defeasible reason for expecting it to be red at t_2, and its looking blue at t_2 provides a defeasible reason for thinking it blue and hence not red at t_2. If these reasons were of the same strength, there would be no basis for preferring one conclusion to the other and the agent would be unable to draw a justified conclusion about the color of the object at t_2. The situation is resolved by noting that the reason for thinking the object is still red at t_2 is weaker than the reason for thinking it was red at t_1, and hence weaker than the reason for thinking the object is blue (and so not red) at t_2. Because the agent has a stronger reason for thinking the object is blue at t_2 than for thinking it is red at t_2, that becomes the justified conclusion and the agent is able to conclude that the object has changed color.

The preceding example illustrates the importance of incorporating an account of degrees of justification into a system of defeasible reasoning. The question addressed in this paper is how the degree of justification of a conclusion should be determined. This is made more complicated by the fact that a conclusion may be supported by several different arguments. The arguments are typically defeasible, and there may also be arguments of varying strengths for defeaters for some of the supporting arguments. What is sought is a way of computing the "on sum" degree of justification of a conclusion. It is clear that three variables, at least, are involved in determing degrees of justification. The reason-strengths of the reason-schemas employed in the argument are relevant. The degrees of justification of the premises are relevant. And the degrees of justification of any defeaters for defeasible steps of the argument are relevant. Other variables might be relevant as well. I am going to assume that reason-strengths and degrees of justification are measurable as extended real numbers (i.e., the reals together with ∞). The justification for this assumption will be provided in section six.

---

[3] The reason-schema proposed in Pollock (1998) involves some additional qualifications, but they are not relevant to the present discussion.

[4] This is discussed in detail in Pollock (1998).

2. Argument-Strengths[5]

Let us begin by looking at arguments for which we have no arguments supporting defeaters. Let the *strength of an argument* be the degree of justification it would confer on its conclusion under those circumstances. A common and seductive view would have it that argument strength can be modeled by the probability calculus. On this view, the strength a conclusion gains from the premises can be computed in accordance with the probability calculus from the strength of the reason (a conditional probability) and the probabilities of the premises. I, and many other authors, have argued against this view at length, but it has a remarkable ability to keep attracting new converts.

There are a number of familiar arguments against the probabilistic model.[6] The simplest argument proceeds by observing that the probabilistic model would make it impossible to be justified in believing a conclusion on the basis of a deductive argument from numerous uncertain premises. This is because as you conjoin premises, if degrees of support work like probabilities, the degree of support decreases. Suppose you have 100 independent premises, each highly probable, having, say, probability .99. According to the probability calculus, the probability of the conjunction will be only .37, so we could never be justified in using these 100 premises conjointly in drawing a single conclusion. But this flies in the face of human practice. For example, an engineer building a bridge will not hesitate to make use of one hundred independent measurements to compute (deduce) the correct size for a girder. I do not have time to discuss this issue at length, so I am just going to assume that deductive arguments provide one way of arriving at new justified conclusions on the basis of earlier ones. A corollary is that the probabilistic model is wrong.

If deductive reasoning automatically carries us from justified conclusions to justified conclusions, then the degree of support a deductive argument confers on its conclusion cannot decrease as the number of premises increases. The degree of justification for the conclusion must be no less than that of the most weakly justified premise. This is the *Weakest Link Principle*, according to which a deductive argument is as good as its weakest link. More precisely:

> The argument strength of a deductive argument is the minimum of the degrees of justification of its premises.

This formulation of the weakest link principle applies only to deductive arguments, but we can use it to obtain an analogous principle for defeasible arguments. If P is a defeasible reason for Q, then we can use conditionalization to construct a simple defeasible argument for the conclusion $(P \supset Q)$, and this argument turns upon no premises:

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

Therefore, $(P \supset Q)$.

As this argument has no premises, the degree of support of its conclusion should be a function of nothing but the strength of the defeasible reason. The next thing to notice is that any defeasible argument can be reformulated so that defeasible reasons are only used in subarguments of this form, and then all subsequent steps of reasoning are deductive. The conclusion of the defeasible argument is thus a deductive consequence of the premises together with a number of conditionals

---

[5] This section and the next summarize arguments proposed in my (1995).

[6] My arguments are given in my (1986), (1995), and in Pollock and Cruz (1999).

justified in this way. By the weakest link principle for deductive arguments, the degree of support of the conclusion should then be the minimum of (1) the degrees of justification of the premises used in the argument and (2) the strengths of the defeasible reasons.

> The argument strength of a defeasible argument is the minimum of the strengths of the defeasible reasons employed in it and the degrees of justification of its premises.

The problem of computing argument strengths is thus computationally simple.

3. The Accrual of Reasons

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

(1) prob(*P* is true / *S* asserts *P*).

Given that this probability is high, I have a defeasible reason for believing that the president of Fredonia has been assassinated if Brown tells me that the president of Fredonia has been assassinated.

In the discussion of the weakest link principle, I urged that argument strengths do not conform to the probability calculus. However, that must be clearly distinguished from the question of whether probabilities license defeasible inferences. In fact, I think that a large proportion of our defeasible inferences are based upon probabilities. Such inferences proceed in terms of one particular reason-schema—the *statistical syllogism*. In my my (1990), I proposed that the principle of the statistical syllogism can be formulated roughly as follows:

(SS1) If $r > 0.5$, then "Fc & prob(G/F) $\geq r$" is a defeasible reason for believing "Gc", the strength of the reason depending upon the value of r.

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

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

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

All examples I have considered that seem initially to illustrate the accrual of reasons turn out in the end to have this same form. They are all cases in which we can estimate probabilities analogous to (2) and make our inferences on the basis of the statistical syllogism rather than on the basis of the original reasons. Accordingly, I doubt that reasons do accrue. It is at least simpler to assume that they do not. If we have two separate undefeated arguments for a conclusion, the degree of justification for the conclusion is the maximum of the strengths of the two arguments. This will be my assumption.

4. The Influence of Defeaters

Thus far I have considered how reason-strengths and the degrees of justification of premises effect the degree of justification of a conclusion. The third variable we must consider is the presence of arguments supporting defeaters. My analysis of the effect of this variable will turn on a taxonomy of defeaters. 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 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 "$(\Pi\Gamma \otimes P)$".

The question arises whether the are any kinds of defeaters other than rebutting and undercutting defeaters. A number of authors have advocated *specificity defeaters*. These have been formulated differently by different authors, but the general idea is that if two arguments lead to conflicting conclusions, but one argument is based upon more information than the other then the "more informed" argument defeats the other. A phenomenon like this is common in reasoning from the statistical syllogism. It can be accommodated by endorsing the following undercutting defeater for (SS1):

> "Hc & $\text{prob}(G/F\&H) \neq \text{prob}(G/F)$" is an undercutting defeater for (SS1).

I refer to these as *subproperty defeaters*. This defeater amount to a kind of "total evidence requirement". It requires us to make our inference on the basis of the most comprehensive facts regarding which we know the requisite probabilities.[7]

Early work in AI on defeasible reasoning tended to concentrate on examples that were

[7] I first pointed out the need for subproperty defeaters in my (1983). Touretzky (1984) subsequently introduced similar defeaters for use in defeasible inheritance hierarchies.

instances of the statistical syllogism (e.g., the venerable "Tweety" arguments), and this led people to suppose that something like subproperty defeat was operative throughout defeasible reasoning. However, I do not see any reason to believe that. To the best of my knowledge, there has never been an intuitive example of specificity defeat presented anywhere in the literature that is not an example of subproperty defeat for a defeasible inference in accordance with the statistical syllogism. Accordingly, I will assume that undercutting defeaters and rebutting defeaters are the only possible kinds of defeaters.

4.1 Diminishers

Now, suppose we have only two arguments to consider, and the conclusion of one of them is a defeater for the final step of the other, as diagrammed in figure 1 (where the "fuzzy" arrow represents defeat). How should this effect the degree of justification of *R*?

It seems clear that if the argument strength of argument #2 is as great as that of argument #1, then the degree of justification of *R* should be 0. But what if the argument strength of argument #2 is less than that of argument #1? In my (1995), I maintained that defeat was an all-or-nothing matter, and hence weaker defeaters leave arguments unaffected. In the scenario just described, this has the consequence that the degree of justification of *R* is the same as the argument strength of argument #1. However, there are some examples that now convince me that this is incorrect. The simplest examples have to do with biased lotteries. To see how these examples work, first consider fair lotteries.

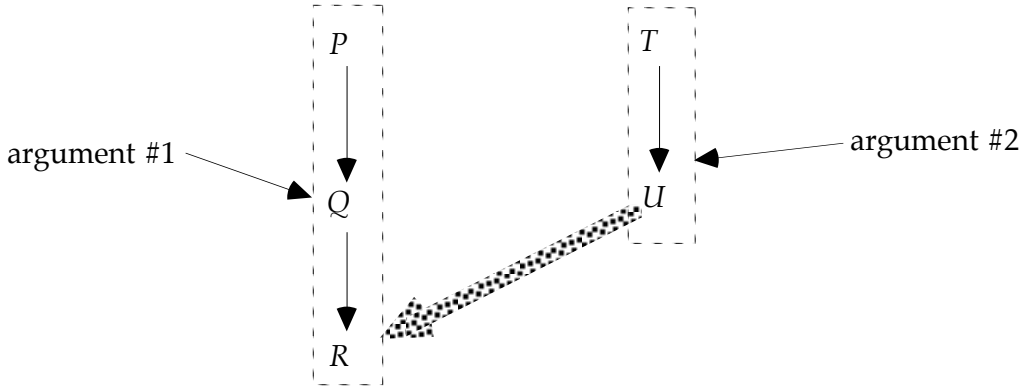

Figure 1. Defeating argument.

Consider a fair lottery consisting of 1 million tickets, and suppose it is known that one and only one ticket will win. Suppose Jones wins the lottery. Observing that the probability is only .000001 of any particular ticket being drawn given that it is a ticket in the lottery, it seems initially reasonable to accept the conclusion regarding any particular ticket that Jones did not hold that ticket. This reasoning is completely general and applies to each ticket. However, these conclusions conflict jointly with something else we are justified in believing, viz., that Jones held some ticket. We cannot be justified in believing each member of an explicitly contradictory set of propositions, and we have no way to choose between them, so it follows intuitively that we are not justified in believing of any ticket that Jones did not hold that ticket.

The formal reconstruction of this reasoning turns on a variant of the statistical syllogism that stands to (SS1) as *modus-tollens* stands to *modus-ponens*:

(SS2) If G is a property that is projectible with respect to a property F and $r > 0.5$, then "$\sim Gc$ & prob(G/F) $\geq r$" is a defeasible reason for believing "$\sim Fc$", the strength of the reason depending upon the value of r.

Let "$T_n x$" be "x has ticket n" and let "Wx" be "x wins". prob($\sim Wx / T_n x$) is high, but upon discovering that Jones won the lottery we would not infer that he did not have ticket n. The reason we would not do so is that this is a case of collective defeat. Because prob($\sim Wx / T_n x$) is high and we know "Wj", we have a defeasible reason for believing "$\sim T_n j$". But for each k, prob($\sim Wx / T_k x$) is equally high, so we have an equally strong defeasible reason for believing each "$\sim T_k j$". From the fact that Jones won, we know that he had at least one ticket. Thus we can construct the counterargument diagramed in figure 2 for the conclusion that "$T_n j$" is true. Our reason for believing each "$\sim T_k j$" is as good as our reason for believing "$\sim T_n j$", so we have as strong a reason for "$T_n j$" as for "$\sim T_n j$". Hence our defeasible reason for the latter is defeated and we are not warranted in believing "$\sim T_n j$".

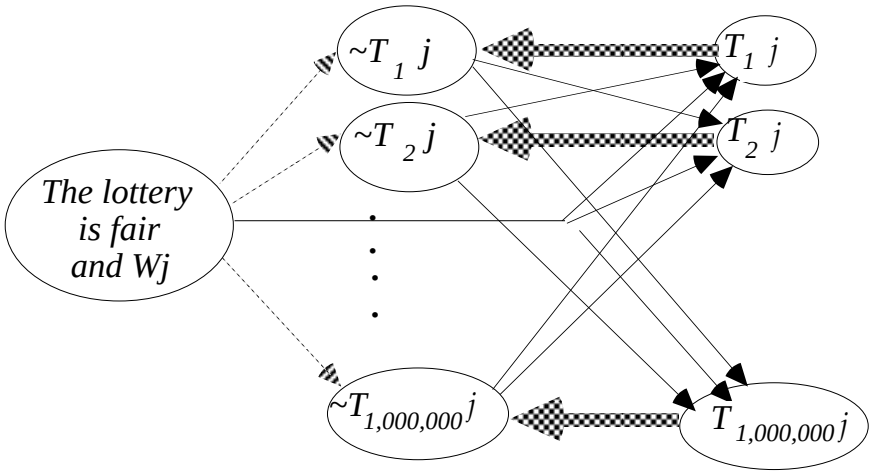

Figure 2. Lottery 1 — a fair lottery

Next, consider lottery 2, which is a biased lottery consisting of just ten tickets. The probability of ticket 1 being drawn is .000001, and the probability of any other ticket being drawn is .111111. It is useful to diagram these probabilities as in figure 3. If Jones wins *this* lottery, it seems reasonable to infer that he did not have ticket 1, because the probability of any other ticket being the winning ticket is more than 100,000 times greater. This inference is supported by (SS2) as follows. As before, for each n, prob($\sim Wx / T_n x$) is fairly high. Combining this with the fact that Jones wins gives us a defeasible reason for believing "$\sim T_n j$", for each n. But these reasons are no longer of equal strength. Because Jones would be much less likely to win if he had ticket 1 than if he had any other ticket, we have a much stronger reason for believing that he does not have ticket 1. As before, for $n > 1$, we have the counterargument diagramed above for "$T_n j$", and that provides as good a reason for believing "$T_n j$" as we have for believing "$\sim T_n j$". Thus the defeasible reason for "$\sim T_n j$" is defeated. But we do not have as good a reason for believing "$T_1 j$" as we do for believing "$\sim T_1 j$". An argument is only as good as its weakest link, and the counterargument for "$T_1 j$" employs the defeasible reasons for "$\sim T_n j$" for $n > 1$. These reasons are based upon lower probabilities (of value .888889) and hence are not as strong as the defeasible reason for "$\sim T_1 j$" (based upon a probability of value .999999). Thus, although we have a reason for believing "$T_1 j$", we have a much better reason for believing "$\sim T_1 j$", and so on sum we are justified in believing the latter.

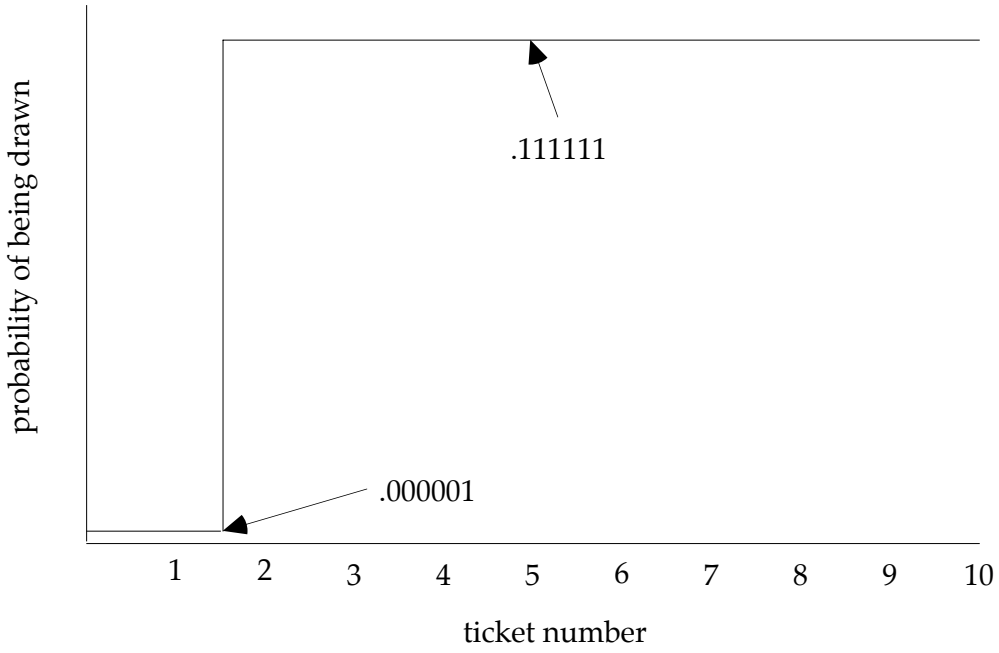

Figure 3. Lottery 2.

Now contrast lottery 2 with lottery 3, which consists of 10,000 tickets. In lottery 3, the probability of ticket 1 being drawn is still .000001, but the probability of any other ticket being drawn is .000011. This is diagramed as in figure 4. If Jones wins lottery 3, it may still be reasonable to infer that he did not have ticket 1, but, and this is the crucial observation, the justification for this conclusion does not seem to be nearly so strong. This is because although we have the same defeasible argument for "~T_1j", the reasons involved in the counterargument for "T_1j" are now much better, being based upon a probability of .999989. They are still not strong enough to defeat the argument for "~T_1j" outright, but they seem to weaken the justification. Thus the degree of justification for "~T_1j" is lower in lottery 3 than it is in lottery 2. The difference between lottery 2 and lottery 3 seems to illustrate that defeaters that are too weak to defeat a conclusion outright may still lower the degree of justification. In other words, they act as *diminishers*.

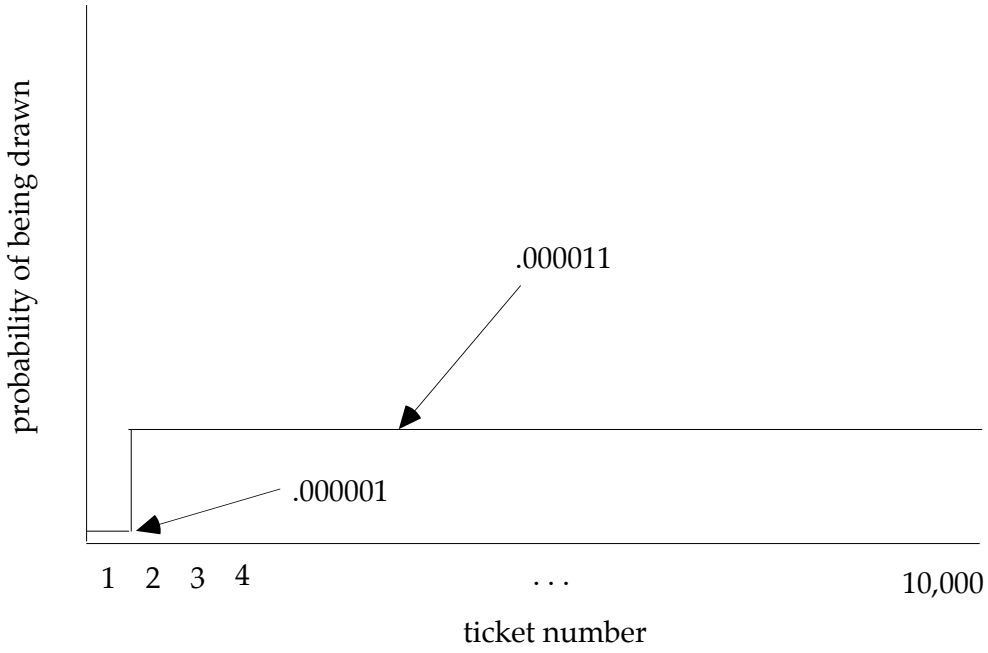

Figure 4. Lottery 3.

Biased lotteries illustrate that given an undefeated argument for P and an otherwise undefeated weaker argument for $\sim P$, the degree of justification for P should be decremented from the argument strength of the first argument by an amount determined by the argument strength of the second argument. That is, there should be a function $\jmath$ such that given two arguments that rebut one another, if their strengths are x and y, the degree of justification for the conclusion of the former is $\jmath(x,y)$. $\jmath$ must satisfy the following conditions:

(3) $\jmath(x,y) \leq x$
(4) $\jmath(x,0) = x$
(5) if $y \geq x$ then $\jmath(x,y) = 0$
(6) if $x \geq z$ and $w \geq y$ then $\jmath(x,y) \geq \jmath(z,w)$.

These cases leave undetermined how $\jmath(x,y)$ and $\jmath(z,w)$ compare in cases in which $x \geq z$ and $w < y$. To resolve these cases, let us make the further assumption that

(7) $\jmath(x+\varepsilon,y+\varepsilon) = \jmath(x,y)$.

This is a "linearity assumption". It tells us that increasing the argument strength and the defeat strength by the same amount ε leaves the resulting degree of justification unchanged. With this further assumption, it becomes determinate how $\jmath(x,y)$ and $\jmath(z,w)$ compare, for any choice of x,y,z,w. This is because (7) implies that $\jmath(x,y) = \jmath(z,y-(x-z))$. Thus $\jmath(x,y) \geq \jmath(z,w)$ iff $\jmath(z,y-(x-z)) \geq \jmath(z,w)$, which by (6) holds iff $w \geq y-(x-z)$, which holds iff $x-y \geq z-w$. Let us define:

$$x \dot{-} y = \begin{cases} x - y & \text{if } y < x \\ 0 & \text{otherwise} \end{cases}$$

Then we have in general:

$\jmath(x,y) \geq \jmath(z,w)$ iff $x \dot{-} y \geq z \dot{-} w$.

So for comparison purposes, we can just as well take $\jmath$ to be $\ominus$. I will thus make the assumption:

> Given an otherwise undefeated argument of strength x supporting P, and an otherwise undefeated argument of strength y supporting $\sim P$, and no other relevant arguments, the degree of justification of P is $x \ominus y$.

4.2 Mixing Rebutting and Undercutting Defeaters

The case of a biased lottery is a case in which the only defeaters are rebutting defeaters. When we turn to more complex examples in which both rebutting and undercutting defeaters are present, things become more complicated. Defeaters for an inference from P to Q are all of two sorts—they are either reasons for $\sim Q$ or reasons for $(P \otimes Q)$. The argument propounded above against the accrual of reasons has the consequence that two arguments supporting the rebutting defeater or two arguments supporting the undercutting defeater provide no stronger support for the defeater than the stronger of the two arguments by itself. However, this leaves open the possibility that the combination of an argument supporting the rebutting defeater and an argument supporting the undercutting defeater might be more potent than either by itself. To test this possibility, consider a case of direct inference. In direct inference we reason from general probabilities (symbolized using "prob") to single case probabilities (symbolized using "PROB"). The basic idea behind classical direct inference was first articulated by Hans Reichenbach (1949): in determining the probability that an individual c has a property F, we find the narrowest reference class X for which we have reliable statistics and then infer that $\mathsf{PROB}(Fc) = \text{prob}(Fx/x \in X)$. For example, insurance rates are calculated in this way. There is almost universal agreement that direct inference is based upon some such principle as this, although there is little agreement about the precise form the theory should take.[8] In my (1990) I proposed reconstructing direct inference as defeasible reasoning that proceeds primarily in terms of the following two principles:

(DI) "$\text{prob}(F/G) = r$ & $\mathbf{J}(Gc)$ & $\mathbf{J}(P \equiv Fc)$" is a defeasible reason for "$\mathsf{PROB}(P) = r$".

(SD) "$\text{prob}(F/H) \neq \text{prob}(F/G)$ & $\mathbf{J}(Hc)$ & $\Box\forall(H \supset G)$" is an undercutting defeater for (DI).

Here "$\mathbf{J}P$" means "P is justified". The defeaters provided by (SD) are also called *subproperty defeaters*. To illustrate, suppose you are an insurance actuary computing auto insurance rates for Helen. You know that Helen is female, and the probability of a female driver having an accident within one year is .03. This gives you a defeasible reason for concluding that the probability of Helen having an accident is .03. But you also know that Helen is a reckless driver, and the probability of a female reckless driver having an accident in a year is .1. This gives you a defeasible reason for the conflicting conclusion that the probability of Helen having an accident is .1. Because the latter inference is based upon more information, you will accept it and reject the former inference. Formally, this is because the latter inference is based upon information that provides a subproperty defeater for the former inference, so the former inference is defeated and the latter is left undefeated. Let us examine this formal reconstruction carefully. There are three arguments involved:

Argument 1 — for the conclusion that the probability of Helen having an accident is .03.

Argument 2 — for the conclusion that the probability of Helen having an accident is .1.

Argument 3 — for the conclusion that Helen is a reckless driver, where the probability of a female driver having an accident is different from the probability of a female reckless driver having an accident.

---

[8] For instance, see Kyburg (1974) and Levi (1977).

Arguments 1 and 2 provide rebutting defeaters for each other. If they are of equal strength and there were no other relevant arguments, then both would be defeated and you would be unable to draw any conclusion about the probability of Helen having an accident. However, the conclusion of argument 3 supports an undercutting defeater for argument 1, so argument 1 is defeated, leaving argument 2 undefeated. The inference graph representing all three arguments can be diagrammed as in figure 5.

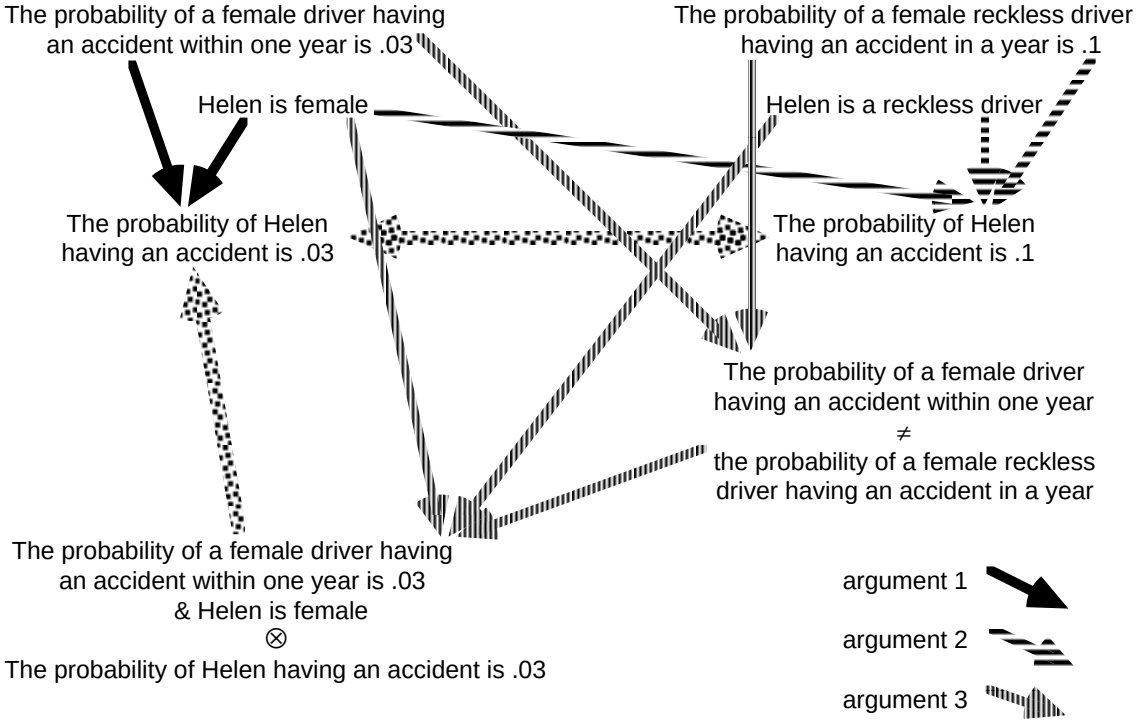

Figure 5. Direct inference

There is, however, a problem with this reconstruction of the reasoning. Recall the prior discussion of temporal projection. Temporal projection provides a defeasible reason for thinking that if something is true at one time then it will still be true later. Some form of temporal projection has been endorsed by most recent authors discussing the frame problem.[9] My formulation of temporal projection differs from most, however, in that the reason-strength decreases as the temporal interval increases. This is the *temporal decay* of temporal projection. I take it that this is fairly intuitive, and it is required to make temporal updating work properly.

Now let us consider how the temporal decay of temporal projection interacts with the preceding example. Imagine you learned that Helen was a reckless driver before you learned that Helen was female. Arguments 2 and 3 both depend upon knowing that Helen is still a reckless driver, which is inferred by temporal projection. Argument 1 depends only upon knowing that Helen is still female, which is more strongly supported. This implies that the argument strengths for arguments 2 and 3 are both less than the argument strength for argument 1. The result is that neither the rebutting defeater provided by argument 2 nor the undercutting defeater provided by argument 3 should be sufficient by itself to defeat argument 1. They would only diminish its strength. On the other hand, argument 2 would be defeated outright by the more strongly

[9] This was proposed by Sandewall (1972), and subsequently endorsed by McDermott (1982), McCarthy (1986), and virtually all subsequent authors.

supported argument 1. This is the intuitively wrong result.

It appears that the only way to get argument 1 defeated and argument 2 undefeated is to allow the rebutting defeater provided by argument 2 and the undercutting defeater provided by argument 3 to work in unison rather than separately. The mechanism I have in mind is that because the undercutting defeater is undefeated, it lowers the degree of support for the conclusion of argument 1. It is that lowered support that should then be compared with the support for the rebutting defeater provided by argument 2, and as the latter is greater, argument 1 is defeated.

Making the proposal more precise, suppose we have an argument of strength x for P, and both rebutting and undercutting defeaters for the final inference. Let r be the strength of the strongest otherwise undefeated rebutting argument and u the strength of the strongest undercutting defeater. If the undercutting defeater is undefeated, it by itself diminishes the strength of the conclusion P to $(x \ominus u)$. That is then compared with r, and if $r > (x \ominus u)$, the inference to P is defeated outright and the inference to $\sim P$ is justified to degree $r \ominus (x \ominus u)$. In general, the justification for P is $(x \ominus u) \ominus r$, and the justification for $\sim P$ is $r \ominus (x \ominus u)$. This amounts to a version of the accrual of defeat, where what accrue are the strongest undercutting defeater and the strongest rebutting defeater.

More generally still, we might also have an undercutting defeater for the argument for $\sim P$. If x is the strength of an argument and u is the strength of its strongest undercutting defeater, let the *discounted argument strength* of the argument be $(x \ominus u)$. Then my provisional proposal can be expressed reasonably simply as follows:

(DJ) Let x be the discounted argument strength of the argument supporting P and r the maximal discounted argument strength of all arguments supporting $\sim P$ (or 0 if there are none). Then:

The degree of justification for P is $(x \ominus r)$

The degree of justification for $\sim P$ is $(y \ominus r)$.

If there are several different arguments supporting P, then the degree of justification of P is the maximal value computed in this way for all the different arguments. In section six, it will turn out that there are some complex examples (involving collective defeat) in which this provisional proposal fails, but (DJ) will motivate the correct treatment of those cases as well.

There are numerous other cases in which the reasoning has an analogous structure, and the temporal decay of temporal projection causes similar prima facie difficulties. In each case, this treatment of defeaters resolves the prima facie difficulties. For example, the Yale Shooting Problem has played an important role in discussions of the Frame Problem.[10] In the Yale Shooting Problem, we are given that a gun is initially loaded. Then it is pointed at Jones and the trigger is pulled. We suppose we know (simplistically) that if a loaded gun is pointed at someone and the trigger pulled, that person will shortly become dead. The conclusion we are supposed to draw in this case is that Jones will die. The Yale Shooting Problem is the problem of showing how this conclusion can be justified. Just as in the case of direct inference, there are two arguments supporting conflicting conclusions:

Argument 1 — for the conclusion that Jones will still be alive after the shooting, based upon temporal projection from the fact that Jones was initially alive.

Argument 2 — for the conclusion that Jones will be dead after the shooting, based upon causal knowledge and the temporal projection that the gun will still be loaded when the trigger is pulled.

---

[10] Hanks and McDermott (1986).

In the absence of further arguments, these arguments will defeat each other, leaving us with no justified conclusion to draw about the state of Jones' health. To get the intuitively correct answer, we need a third argument:

Argument 3 — supporting an undercutting defeater for argument 1.

The Yale Shooting Problem is resolved by explaining the details of argument 3. I have proposed such a solution in my (1998). For present purposes, the details are not important. Suffice it to say that the undercutter turns upon causal knowledge and the premise that the gun is still loaded when the trigger is pulled.

If we ignore the temporal decay of temporal projection, the foregoing constitutes a solution to the Yale Shooting Problem. But now suppose we observe the gun to be loaded *before* observing Jones to be alive. In this case the strengths of arguments 2 and 3, depending as they do on inferring that the gun is still loaded when the trigger is pulled, may both be less than the strength of argument 1.

The temporal profile we should get is the following. If we observe the gun to be loaded long before observing Jones to be alive (e.g., years ago), our justification for believing Jones to remain alive should be weakened but not defeated. On the other hand, if we observe the gun to be loaded just shortly before observing Jones to be alive, we should be able to conclude that Jones will die. The intuitive rationale for this profile seems to be as follows. First, we have the undefeated argument 3 for the undercutting defeater for argument 1, but it is weaker than argument 1. Instead of defeating argument 1 outright, it weakens it seriously. This leaves us with only a weak reason for thinking that Jones remains alive. Then argument 2 provides a reason for thinking Jones is dead. If argument 2 turns upon a temporal projection from the distant past that the gun is loaded, it will not be strong enough to defeat even the weakened argument 1, but if the temporal projection is from a recent observation then argument 2 will be strong enough to defeat the weakened argument 1. This is exactly the computation proposed by principle (DJ).

Both of these examples turn upon the temporal decay of temporal projection. However, we can contrive other examples illustrating principle (DJ) that do not. For example, the strength of the reason provided by the statistical syllogism depends upon the value of the probability. Consider a variant of the Yale Shooting Problem in which the gun is equipped with a red light that is supposed to be on when the gun is loaded and off when the gun is unloaded. This mechanism is highly reliable, but like any mechanical contrivance, it is not totally reliable. Thus we are reasoning in accordance with the statistical syllogism when we judge the gun to be loaded because the light is on. We can consider variants of the example in which the warning mechanism varies in reliability. If we adjust the reliability so that the statisical syllogism provides only as a strong a reason as a temporal projection from the distant past, then we would not be able to conclude that Jones will die as a result of our pulling the trigger, but if we adjust the reliability so that the reason is as strong as a temporal projection from the immediate past, then we should be able to conclude that Jones will die. Again, to get this result we must let rebutting defeaters and undercutting defeaters act in unison. So the logic of this example does not depend upon the use of temporal projection.

The upshot is that to get the intuitively correct answer in cases involving simultaneous rebutting and undercutting defeat, we must allow the two defeaters to work in unison. This amounts to a limited version of the accrual of defeat.

5. Computing Defeat Statuses

The problem discussed in section four is that of computing the degree of justification for the conclusion of an argument when we already know the degrees of justification for the premises of the arguments and for any relevant undercutting defeaters. That, however, is a somewhat

artificial problem. The more general problem that must be addressed is that of computing degrees of justification for the conclusions of each member of a set of interrelated arguments, some of which support defeaters for others. In the general problem, we cannot make any assumptions about the degrees of justification for the premises and defeaters for most of the arguments.

It is natural to suppose that the general problem can be solved by applying principle (DJ) recursively. We begin with a set of "initial premises" that are given in some fashion (perhaps by perception) and for which there are no defeating arguments. For the initial premises, the degree of justification is simply given. Then we apply (DJ) recursively to compute the degrees of justification for all other conclusions.

Unfortunately, matters are not so simple. The proposed recursion assumes that the set of arguments can be ordered in such a way that each argument comes *after* all arguments for its premises and all arguments for undercutting defeaters for its steps. Without this *linearity assumption*, there is no reason to expect the recursive computation to assign degrees of justification to every conclusion. And, unfortunately, the linearity assumption need not hold. Arguments can form a tangled web in which no such linearization is possible. To consider a very simple example, suppose Brown tells me that Smith is an unreliable informant, and Smith in turn tells me that Brown is an unreliable informant. Because we know that people are generally reliable informants, a person's telling us something gives us a reason (in accordance with the statistical syllogism) for believing what we are told. Thus we have the two interrelated arguments of figure 6. The problem is that each supports an undercutting defeater for the other, so there is no way to order them so that each one comes after all arguments supporting undercutting defeaters for it, and accordingly there is no way to compute degrees of justification recursively.

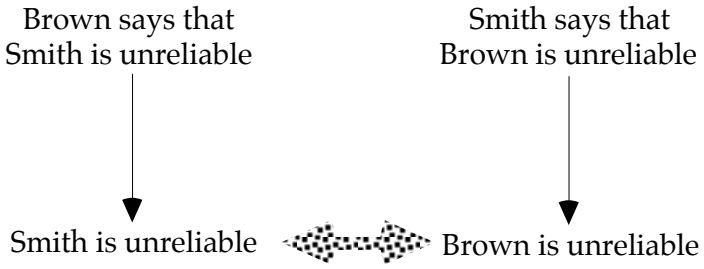

Figure 6. Mutual undercutting defeat

A solution to the general problem of computing degrees of justification must incorporate a solution to the problem of determining which arguments are defeated and which are undefeated. These are not really separable problems, because in accordance with principle (DJ), degrees of justification are a function in part of defeat statuses (the rebutting and undercutting arguments must be "otherwise undefeated"), but defeat statuses are also a function of degrees of justification—an argument is defeated just in case the degree of justification it confers on its conclusion is 0. These two problems can be unified by identifying the degree of justification of a conclusion with the defeat status of the strongest undefeated argument supporting it. In other words, defeat statuses are not just "defeated" and "undefeated". Defeat statuses are numbers corresponding to degrees of justification, and a defeated argument is one that is completely devoid of justification, i.e., one for which the defeat status is 0.

The problem of computing degrees of justification then becomes the problem of computing defeat statuses. A theory of defeasible reasoning must provide an analysis of defeat status when we are presented with a set of arguments some of which support defeaters for others. The proposal that I made in my (1994) and (1995) was to define *status assignments* to be assignments of defeat status to the conclusions of arguments, where the assignments are required to be consistent with the rules I had previously defended regarding degrees of justification. The result was the analysis given in section one. Note, however, that the rules used there for assigning defeat statuses within a status assignment presuppose that defeaters do not accrue. For an

argument to be defeated relative to an assignment, there must be a single defeater that is capable of defeating it. These rules must be modified to accommodate principle (DJ). We must allow undercutting and rebutting defeaters to work in concert, and to do that the status assignment must keep track of degrees of justification (relative to the assignment)—not just the defeat statuses "defeated" and "undefeated". A node will be defeated in an assignment just in case the assignment assigns 0 to it.

The plan is to redefine a status-assignment to be an assignment of defeat-statuses consistent with (DJ). Principle (DJ) was formulated as follows:

(DJ) Let x be the discounted argument strength of the argument supporting P and r the maximal discounted argument strength of all arguments supporting $\sim P$ (or 0 if there are none). Then:

The degree of justification for P is $(x \ominus r)$

The degree of justification for $\sim P$ is $(y \ominus r)$.

In translating this into a rule for computing defeat-statuses, we must reflect upon what x and r are supposed to be. The argument supporting P corresponds to a node of the inference graph, but x is not the same thing as the defeat-status of that node. Rather, x is supposed to be the strength the argument would confer on its conclusion *in the absence of rebutting defeaters*. Let us call this the *discounted argument strength* an assignment assigns to a node, and define it precisely as follows:

A partial status assignment σ assigns an *discounted argument strength* v to a node n iff σ assigns values to every member of the node-basis of n and to every undercutting defeater for n, and where β is the minimum of the values σ assigns to the members of the node-basis and the reason-strength of the reason supporting n, and δ is the maximum value σ assigns to an undercutting defeater of n (or 0 if there are none), then $v = \beta \ominus \delta$.

So x should be the discounted argument strength of the node. Similarly, r should be the maximal discounted argument strength of any rebutting node. Then we can incorporate (DJ) into the defeat-status computation by defining:

σ is a *partial status assignment* iff σ is a function assigning an extended real number to a subset of the nodes of an inference-graph in such a way that:

1. σ assigns its strength to any initial node;[11]
2. If σ assigns 0 to some member of the node-basis of a node then σ assigns 0 to the node;
3. If σ assigns a value to either a node n or some member of the node basis of n that is less than or equal to the value it assigns to some defeater for n, then σ assigns 0 to n;
4. If σ assigns values to an undercutting defeater u, a rebutting defeater r, and a member of the basis b of a node n in such a way that $\sigma(b) \leq \sigma(r) + \sigma(u)$, then σ assigns 0 to n;
5. If σ assigns values to a node n, an undercutting defeater u, and a rebutting defeater r in such a way that $\sigma(n) \leq \sigma(r) + \sigma(u)$, then σ assigns 0 to n;
6. If σ assigns a non-zero value to a node n, σ assigns values to all members of the node-basis of n and all defeaters for n;
7. If σ assigns values to every member of the node-basis of a node n, and σ assigns values to every defeater for n, let δ_u be the maximum value σ assigns to any undercutting defeaters for n (or 0 if there are none), let δ_r be the maximum discounted argument strength σ assigns to any rebutting defeaters for n (or 0 if there are none), and let β be the minimum

[11] It is assumed that initial nodes are assigned strengths directly.

of the values σ assigns to the members of the node-basis and the reason-strength of the reason supporting *n*. Then σ assigns $\beta \ominus (\delta_u + \delta_r)$ to *n*.

My proposal is then:

A node is undefeated iff every status assignment assigns a non-zero value to it; otherwise it is defeated.

The *undefeated strength* of an argument is the minimal value *r* such that there is a status assignment assigning *r* to the node representing the argument.

The degree of justification of a conclusion is the maximum undefeated strength of all arguments supporting it.

To illustrate this analysis, consider the simple case of collective rebutting defeat diagrammed in figure 7. Suppose the initial nodes *A* and *B* are assigned strength 1. Let the defeat-statuses assigned to *B* and ~*B* be *x* and*y* respectively. As there are no undercutting defeaters, the discounted argument strength of *B* is 1 and the discounted argument strength of ~*B* is 1. If $x \leq y$ then by (4), $x = 0$, and hence by (6) $y = 1 - x = 1$. If $x > y$ then by (4), $y = 0$ and hence $x = 1 - y = 1$. So there are two possible assignments, with *B* defeated in one and ~*B* defeated in the other. Then, as expected in a case of collective defeat, both nodes are defeated and the degree of justification for both *B* and ~*B* is zero.

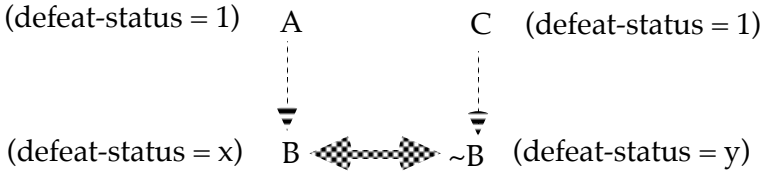

Figure 7. Collective rebutting defeat

It is instructive to contrast collective undercutting defeat, wherein two nodes provide undercutting defeaters for each other, with the case of collective rebutting defeat diagrammed in figure 7. A simple example of collective undercutting defeat was diagrammed in figure 6. For simplicity, suppose the initial nodes are assigned strength 1. Suppose "Smith is unreliable" is assigned strength *x*, and "Brown is unreliable" is assigned strength *y*. Then the discounted argument strengths of these nodes are 1–*y* and 1–*x*, respectively. So by clause (vii) of the definition of "partial status-assignment", $x = 1-y$ and $y = 1-x$, i.e., $x+y = 1$. There are infinitely many ways of assigning values to *x* and *y* to satisfy this constraint. However, clause (iii) requires that if $x \geq y$ then $y = 0$ and hence $x = 1$, and if $y \geq x$ then $x = 0$ and hence $y = 1$. So there are just two status-assignments.

In cases of collective defeat, there are multiple status-assignments. This generates a distinction between two ways a node can be defeated. A node is *defeated outright* iff every status-assignment assigns 0 to it. A node is *provisionally defeated* iff some status-assignment assigns 0 to it but some other status-assignment assigns a non-zero value to it. The significance of provisional defeat is that provisionally defeated nodes, although defeated, retain the ability to defeat other nodes. This is because defeatees of a provisionally defeated nodes can be defeated in any status-assignment in which the provisionally defeated node is assigned a nonzero value.

Provisional defeat prevents this analysis from validating (DJ). However, where there is a difference, the present account seems to be correct. To illustrate, consider the inference-graph diagrammed in figure 8. Suppose Smith and Brown accuse each other of being unreliable. Then we have a case of collective undercutting defeat, and should remain agnostic about whether either is reliable. Under these circumstances, if Smith also tells us that it is raining, it would be unreasonable to take his word for it. This illustrates that the conclusion that Smith is unreliable

can defeat the inference from Smith's saying that it is raining to its raining, even though the conclusion that Smith is unreliable is defeated. What is crucial to this example is that the defeat is a case of provisional defeat. The analysis proposed above gives the right answer in this example. Principle (DJ), on the other hand, does not. Because the only defeater for "It is raining" is defeated, (DJ) would have "It is raining" undefeated.

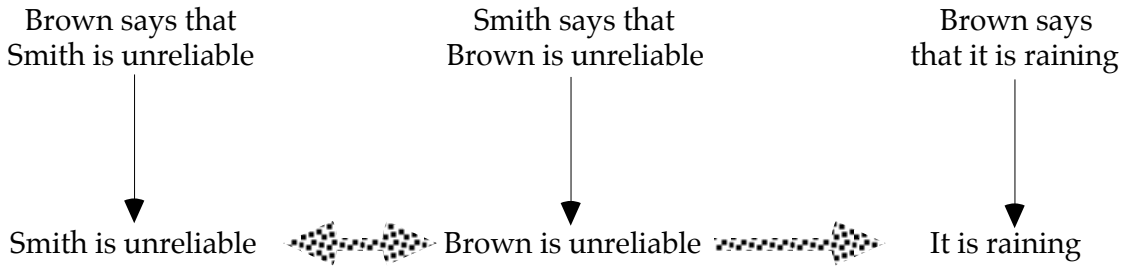

Figure 8. Defeat by provisionally defeated conclusion

6. Measuring Strengths

It has been assumed throughout that reason-strengths and degrees of justification can be measured using the extended reals, although nothing has been said about how that is done. If we are to take strength seriously, we must have some way of measuring it. One way is to compare reasons with a set of standard equally good reasons that have numerical values associated with them in some determinant way. I propose to do that by taking the set of standard reasons to consist of instances of the statistical syllogism (SS1). For any proposition p, we can construct a standardized argument for $\neg p$ on the basis of the pair of suppositions "prob(F/G) $\geq r$ & Gc" and "($p \equiv \sim Fc$)":

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

where the strength of the argument is a function of r. We can measure the strength of a defeasible reason for p in terms of that value of r such that the conflicting argument from the suppositions "prob(F/G) $\geq r$ & Gc" and "($p \equiv \sim Fc$)" exactly counteracts it. The value r *determines* the reason-strength in the sense that the reason-strength is some function $\mathbf{j}(r)$ of r. It is tempting to identify $\mathbf{j}(r)$ with r, but that may not work. The difficulty is that by identifying $\mathcal{J}$ with $\ominus$, we have required reason-strength to be a cardinal measure that can be meaningfully added and subtracted. Adding and subtracting reason-strengths may not be the same thing as adding and subtracting the corresponding probabilities. For example, should it turn out that the reason-strength corresponding to a probability r is given by $\log(r)$, then adding reason-strengths would be equivalent to multiplying probabilities rather than adding them.

I do not have an a priori argument to offer regarding what function $\mathbf{j}(r)$ produces the reason-strength corresponding to r. The only way to determine this is to look for proposals that work plausibly in concrete examples. Perhaps the most illuminating example is that of the biased lotteries diagrammed in figures 3 and 4. Suppose reason-strengths could be identified with the corresponding probabilities. In lottery 2, the probability of ticket 1 being drawn is .000001, and the probability of any other ticket being drawn is .111111. We wanted to conclude in this case that we are justified in believing that ticket 1 will not be drawn. The probability corresponding to the argument-strength for this conclusion is .999999, however the probability corresponding

to the counter-argument for the conclusion that ticket 1 will be drawn (because no other ticket will) is .888889. The difference between these probabilities is .11111, which is a very low probability. If probabilities and degrees of justification could be identified, i.e., $\mathbf{j}(r) = r$, this would produce too low a degree of justification for it to be reasonable to believe that ticket 1 will not be drawn. So apparently we cannot compute degrees of justification by adding and subtracting probabilities.

There is statistical lore suggesting that in probabilistic reasoning degrees of justification can be compared in terms of likelihood ratios.[12] When (as in the biased lotteries) we have an argument for P based on a probability r, and a weaker argument for $\sim P$ based on a probability r^*, the likelihood ratio is $(1 - r)/(1 - r^*)$. The suggestion is that the degree of justification for $\sim P$ is determined by the likelihood ratio. For example, in lottery 2 the likelihood ratio is .000009, while in lottery 3 it is .09. Note that likelihod ratios are defined so that higher likelihood ratios correspond to lower degrees of justification. An equivalent but more intuitive way of measuring degrees of justification is by using the inverse of the likelihood ratios.

In my (1990) I argued that likelihood ratios seem to yield the intuitively correct answers in many cases of statistical and inductive reasoning, and on that basis I am prepared to tentatively endorse their use in measuring degrees of justification. If we take the degree of justification $\mathbf{j}(r)$ resulting from an application of the statistical syllogism with probability r to be $\log(.5) - \log(1 - r)$, then the result of subtracting degrees of justification is the same as taking the logarithm of the inverse of the liikelihood ratio, i.e., $\mathbf{j}(r) - \mathbf{j}(r^*) = \log(1 - r^*) - \log(1 - r) = \log((1 - r^*)/(1 - r))$. Thus this choice of $\mathbf{j}(r)$ yields congenial results. This will be my proposal:

> If X is a defeasible reason for p, the strength of this reason is $\log(.5) - \log(1 - r)$ where r is that real number such that an argument for $\neg p$ based upon the suppositions "$\text{prob}(F/G) \geq r$ & Gc" and "$(p \equiv \sim Fc)$" and employing the statistical syllogism exactly counteracts the argument for p based upon the supposition X.

The reason for the "log(.5)" term is that this produces the result that $\mathbf{j}(.5) = 0$. An application of the statistical syllogism based on a probability of .5 should produce no justification for the conclusion, and this is captured by setting the degree of justification to be 0. On the other hand, note that $\mathbf{j}(1) = \infty$. That is, the strongest reasons have infinite reason-strength. This could create problems if we ever wanted to subtract the strengths of such arguments from each other, because $\infty - \infty$ is undefined, but in fact we will never have occasion to do that.

7. And/Or Inference-Graphs

In the interest of theoretical clarity, inference-graphs were defined in such a way that different arguments for the same conclusion are represented by different nodes. This made it clearer how the algorithm for computing defeat-status works. However, for the purpose of implementing defeasible reasoning, using different nodes to represent different arguments for the same conclusion is an inefficient representation, because it leads to needless duplication. If we have two arguments supporting a single conclusion, then any further reasoning from that conclusion will generate two different nodes. If we have two arguments for each of two conclusions, and another inference proceeds from those two conclusions, the latter will have to be represented by four different nodes in the inference-graph, and so on. This is illustrated in figure 9, where P and Q are each inferred in two separate ways, and then R is inferred from P and Q.

A more efficient representation of reasoning would take the inference-graph to be an and/or graph rather than a standard graph. In an and/or graph, nodes are linked to *sets* of nodes rather

---

[12] This is known as the likelihood principle. It is due to R. A. Fisher (1922), and versions of it have been endorsed by a variety of authors, including G. A. Barnard (1949 and 1966), Alan Birnbaum (1962), A. W. F. Edwards (1972), and Ian Hacking (1965).

than individual nodes. This is represented diagramatically by connecting the links with arcs. In an and/or inference-graph, when we have multiple arguments for a conclusion, the single node representing that conclusion will be tied to different bases by separate groups of links. This is illustrated in figure 10 by an and/or inference-graph encoding the same reasoning as the standard inference-graph in figure 9. In an and/or inference-graph, a *support-link* will be a *set* of supporting-arrows connected with an arc.

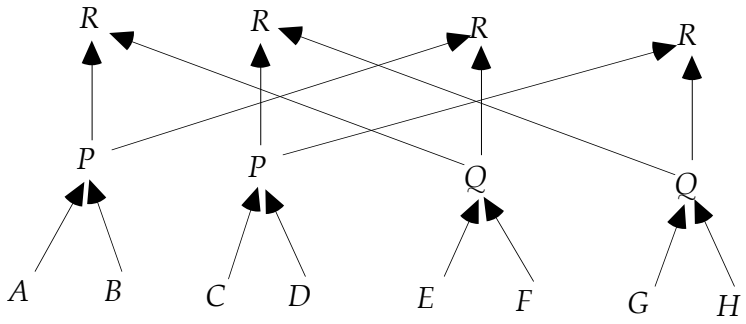

Figure 9. Inference-graph with mutliple arguments for a single conclusion.

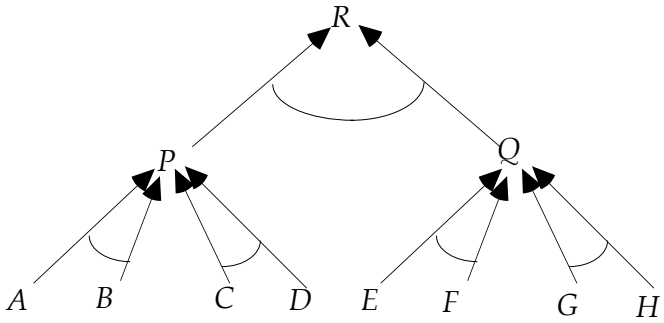

Figure 10. An and/or inference-graph

Although and/or graphs provide an efficient representation of reasoning, they complicate the computation of defeat-status. Using simple inference-graphs, we can use the defeat-status computation of section five to compute defeat-statuses. If we are to use and/or graphs in the implementation, we must find a defeat-status computation for and/or graphs that is equivalent to that for simple inference-graphs. A simple inference-graph can be rewritten as an and/or graph in which each node of the and/or graph corresponds to a set of nodes of the simple graph. A node of the simple inference-graph, on the other hand, corresponds to an *argument* in the and/or graph. An argument is a kind of connected sub-tree of the graph. More precisely:

> An *argument from* an and/or inference-graph G *for* a node N is a minimal subset A of the nodes and support-links of the graph such that (1) if a node in A has any support-links in G, exactly one of them is in A, (2) if a support-link is in A then the nodes in its support-link-basis are also in A, and (3) N is in A.

Nodes in the simple inference-graph correspond one-one to arguments in the and/or inference-graph.

The result we want to ensure is the following *Correspondence Principle*:

A node of the and/or inference-graph is undefeated iff one of the corresponding nodes of the simple inference-graph is undefeated.

The key to achieving this is to have status assignments assign values both to nodes and to support-links. The assignments to support-links for a node will be analogous to assignments to individual nodes of a simple inference-graph that support the same conclusion. It is convenient to define the *discounted argument strength* an assignment assigns to a support-link to be the value it would assign to the link in the absence of rebutting defeaters:

A partial status assignment σ assigns a *discounted argument strength* v to a support-link L iff σ assigns values to every member of the node-basis of L and to every undercutting defeater for L, and where β is the minimum of the values σ assigns to the members of the node-basis and the reason-strength of L, and δ is the maximum value σ assigns to an undercutting defeater of L (or 0 if there are none), then $v = \beta \ominus \delta$.

This is analogous to the discounted argument strength of a node relative to an assignment to a simple inference-graph. Then we can define:

σ is a *partial status assignment* iff σ is a function assigning an extended real number to a subset of the nodes and support-links of an inference-graph in such a way that:

1. σ assigns its strength to any nondoxastic node;
2. If σ assigns 0 to some member of the node-basis of a link then σ assigns 0 to the link;
3. If σ assigns a value to either a link L some member of the node basis of L that is less than or equal to the value it assigns to some defeater for L, then σ assigns 0 to L;
4. If σ assigns values to an undercutting defeater u, a rebutting defeater r, and a member of the basis b of a link L in such a way that $\sigma(b) \leq \sigma(r) + \sigma(u)$, then σ assigns 0 to L;
5. If σ assigns values to a link L, an undercutting defeater u, and a rebutting defeater r in such a way that $\sigma(L) \leq \sigma(r) + \sigma(u)$, then σ assigns 0 to L;
6. If σ assigns a non-zero value to a link L, it assigns values to all members of the node-basis of L and all defeaters for L;
7. If σ assigns values to every member of the node-basis of a link L, and σ assigns values to every defeater for L, let δ_u be the maximum value σ assigns to any undercutting defeaters for L (or 0 if there are none), let δ_r be the maximum discounted argument strength σ assigns to any rebutting defeaters for L (or 0 if there are none), and let β be the minimum of the values σ assigns to the members of the node-basis and the reason-strength of L. Then σ assigns $\beta \ominus (\delta_u + \delta_r)$ to L;
8. If every support-link of a node is assigned 0, the node is assigned 0;
9. If some support-link of a node is assigned a value greater than 0, the node is assigned the maximum of the values assigned to its support-links;
10. If every support-link of a node that is assigned a value is assigned 0, but some support-link of the node is not assigned a value, then the node is not assigned a value.

It is tempting to define:

A node of the and/or inference-graph is undefeated iff every status assignment assigns a non-zero value to it; otherwise it is defeated.

However, this does not make the Correspondence Principle true. Figure 11 is a simple counterexample, with the simple inference-graph on the right and the corresponding and/or graph on the left. In the simple graph, there are two status-assignments, and one assigns "undefeated" to the left "$E \vee F$" and the other assigns "undefeated" to the right "$E \vee F$", but neither "$E \vee F$" is assigned "undefeated" by both status-assignments, so both are defeated. In the and/or graph,

there are also two status-assignments, and each assigns "undefeated" to "$E \vee F$" (by making a different argument undefeated). Thus on the above proposal, "$E \vee F$" would be undefeated in the and/or inference-graph, but defeated in the simple inference-graph.

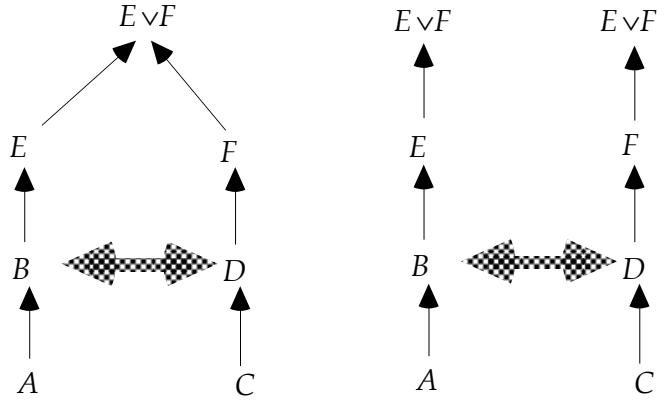

Figure 11. Counterexample to the Correspondence Principle

In figure 11, the difficulty lies in the fact that in the and/or graph there is no argument that comes out undefeated on both assignments. A node's being assigned a value by an assignment means that *there is* an argument supporting the node that has that strength in the assignment. So a node being assigned a nonzero value in every assignment means that in every assignment there is an argument of nonzero strength supporting it, but this does not imply that there is a single argument supporting it that has a nonzero strength in every assignment. The latter is what is required for justified belief. This indicates that the correct analysis of defeat-status for and/or inference-graphs is:

> A node of the and/or inference-graph is undefeated iff the graph contains an argument for the node which is such that every status assignment assigns a non-zero value to all nodes and links in the argument; otherwise it is defeated.

With this definition it becomes simple to prove the Correspondence Principle by induction on the length of the arguments.

What remains is to determine how the degree of justification of a node is determined. It should be the maximum degree of support provided by undefeated arguments for the node. Arguments correspond to nodes of the simple inference-graph, so to duplicate the computation the works for simple inference-graphs, we can compute an argument-strength for an argument in each status-assignment for the and/or graph, define the argument-strength simpliciter to be the minimum of the argument-strengths the argument has in different status-assignments, and then define the degree of justification of a node of the and/or graph to be the maximum argument-strength of any argument for that node. The computation of the argument-strength for an argument in a status-assignment should be parallel to the computation of status assignments in a simple inference-graph. The final step of the argument applies a reason-schema to the conclusions of one or more shorter arguments. Let us call those the "penultimate subsidiary arguments" of the argument. The *presumptive argument strength* of the argument is the minimum of the reason-strength of the last step and the argument-strengths of the penultimate subsidiary arguments. This is the strength the argument would be assigned in the absence of any defeaters. From the presumptive argument strength of the argument we want to substract the sum of (1) the argument strength of the strongest rebutting argument, and (2) the argument strength of the strongest undercutting argument. But (1) and (2) are precisely the values supplied by the status assignment to the sole node of the and/or graph encoding the rebutting defeater and the sole node of the and/or graph encoding undercutting defeater. Thus we can compute argument

strength relative to a status assignment recursively. Recall that an argument is a set of nodes and support-links:

> An *argument from* an and/or inference-graph G *for* a node N is a minimal subset A of the nodes and support-links of the graph such that (1) if a node in A has any support-links in G, exactly one of them is in A, (2) if a support-link is in A then the nodes in its support-link-basis are also in A, and (3) N is in A.

By requiring that arguments be minimal, we ensure that each node in the argument is supported by a single link, and hence by induction, each node is either initial or supported by a single subargument of the given argument. Where σ is a status assignment for G, we define:

(i) if A consists of a single initial node, its argument-strength relative to σ is the strength of the node;

(ii) if L is the final link of A, r is its reason-strength, $b_1,...,b_n$ are the nodes constituting the basis of L, $A_1,...,A_n$ are the arguments for for $b_1,...,b_n$, $r_1,...,r_n$ are the argument strengths of $A_1,...,A_n$ relative to σ, δ_u is the value assigned by σ to the undercutting defeater for L (or zero if G does not contain the undercutting defeater), and δ_r is the value assigned by σ to the rebutting defeater for L (or zero if G does not contain the rebutting defeater), then the argument strength for A relative to σ is $\min(r,r_1,...,r_n) \ominus (\delta_u + \delta_r)$.

> The argument-strength of an argument from an and/or inference-graph G is its minimum argument-strength relative to an assignment.

> The degree of justification of a node in an and/or inference-graph G is the maximum argument-strength of any argument from G for the node.

This produces the same degree of justification for a conclusion as the computation described earlier for simple inference-graphs.

8. Conclusions

In conclusion, I have endorsed a limited version of the accrual of defeat wherein the strongest rebutting defeater and the strongest undercutting defeater for a conclusion can work in unison. I have proposed an analysis of defeat-status based upon this principle. Work is currently underway to implement this new defeat-status computation in OSCAR.

The topic of degrees of justification resulting from defeasible reasoning is virtually unexplored in AI. There is a massive literature on degrees of probability, but this paper partly argues and partly assumes (referring to arguments given elsewhere) that degrees of justification are not probabilities, in the sense that they do not conform to the probability calculus.

References

Barnard, G. A.
1949 Statistical inference. *Journal of the Royal Statistical Society B, II*, 115-149.
1966 The use of the likelihood function in statistical practice. *Proceedings v Berkeley Symposium on Mathematical Statistics and Probability I*, 27-40.
Birnbaum, Allan
1962 On the foundations of statistical inference. *Journal of the American Statistical Association 57*, 269-326.
Edwards, A. W. F.
1972 *Likelihood*. Cambridge: Cambridge University Press.
Fisher, R. A.
1922 On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, *222*, 309-368.
Hacking, Ian
1965 *Logic of Statistical Inference*. Cambridge: Cambridge University Press.
Hanks, Steve, and McDermott, Drew
1986 "Default reasoning, nonmonotonic logics, and the frame problem", AAAI-86.
Kyburg, Henry
1974 *The Logical Foundations of Statistical Inference*. Dordrecht: Reidel.
Levi, Isaac
1977 "Direct inference". *Journal of Philosophy* **74**, 5-29.
McCarthy, John
1986 "Applications of circumscription to formalizing common sense knowledge." *Artificial Intelligence* **26**, 89-116.
McDermott, Drew
1982 "A temporal logic for reasoning about processes and plans", *Cognitive Science* **6**, 101-155.
Pollock, John
1974 *Knowledge and Justification*, Princeton University Press.
1983 "Epistemology and probability", *Synthese* **55**, 231-252.
1987 *Contemporary Theories of Knowledge*, Rowman and Littlefield.
1990 *Nomic Probability and the Foundations of Induction*, Oxford University Press.
1994 "Justification and defeat", *Artificial Intelligence* **67**, 377-408.
1995 *Cognitive Carpentry*, MIT Press.
1997 "Reasoning about change and persistence: a solution to the frame problem", *Nous* **31**, 143-169.
1998 "Perceiving and reasoning about a changing world", *Computational Intelligence*, *Intelligence* **14**, 498-562.
1998a "Degrees of Justification", in P. Weingartner, G. Schurz and G. Dorn (Eds.), *The Role of Pragmatics in Contemporary Philosophy. (Proceedings of the 20th International Wittgenstein Symposium 1997 Kirchberg/Wechsel, Austria)*, Hoelder-Pichler Tempskypublishers, Vienna, 207-223.
1999 "The logical foundations of goal-regression planning in autonomous agents", *Artificial Intelligence*, forthcoming in 1999.
Praaken, H. and G.A.W. Vreeswijk
19?? Logics for Defeasible Argumentation, to appear in *Handbook of Philosophical Logic, 2nd Edition*, ed. D. Gabbay. Dordrecht: Kluwer Academic Publishers.
Reichenbach, Hans
1949 *A Theory of Probability*. Berkeley: University of California Press. (Original German

edition 1935).
Sandewall, Erik
1972 "An approach to the frame problem and its implementation". In B. Metzer & D. Michie (eds.), *Machine Intelligence* **7**. Edinburgh: Edinburgh University Press.
Touretzky, David
1984 "Implicit orderings of defaults in inheritance systems", *Proceedings of AAAI-84.*