May 15th, 9:00 AM - May 17th, 5:00 PM

John L. Pollock's theory of rationality

David Hitchcock
McMaster University

Title: John L. Pollock's theory of rationality
Author: David Hitchcock
Response to this paper is unavailable.

John L. Pollock is trying to build a person.

Specifically, he has constructed a rational architecture, dubbed OSCAR (1999), for an artificial intellect, a system which will take in perceptual input from its surroundings, form beliefs, have likes and dislikes about the situation it believes itself to be in, formulate plans of action for making its situation more to its liking, evaluate and adopt such plans, and carry them out. Such a system, Pollock believes, will have thoughts and feelings in much the same way as human beings do, and thus will have the same moral standing as a human being. In short, it will be a person.

The failure so far of the research program of strong artificial intelligence (strong AI) to fulfill its extravagant promises Pollock attributes to the lack of a computationally implemented general theory of rationality. Researchers in AI, he holds, have relied on unformulated intuitions about rationality, which are not good enough. Building on more than 30 years of research in theoretical and applied epistemology, Pollock proposes to remedy the lack he perceives.

The theory of rationality implemented in OSCAR takes positions on many issues discussed in critical thinking, informal logic and the theory of argumentation. Without accepting (or rejecting) the project of strong AI, I propose to discuss the positions Pollock takes, and the reasons he gives for them. In general, I will be endorsing his arguments. Pollock's theory of rationality, I will argue, is a sophisticated system which anybody with a scholarly interest in what it means to be a rational agent must take seriously. Pollock has invited researchers with interests in such topics as induction, analogy and practical reasoning to use OSCAR as the basis for computational implementation of proposed systems for such reasoning. I believe that we should take him up on this offer.

1. The basis of norms of rationality

Pollock (1995: 1-3) takes a naturalistic approach to the theory of knowledge. He takes it that human beings know how to reason, although they do not always do so correctly and they are in general not introspectively aware of their procedural knowledge. Philosophical epistemology is an attempt to articulate this implicit knowhow, and is thus a way of doing empirical human psychology.

One could object that this approach relativizes the theory of rationality unduly to the accidents of human psychology as it has evolved through the last two million years. Consider a proposed norm of rationality: if a rational agent

believes that a disjunction is true and that one of its two disjuncts is false, that agent should either add the remaining disjunct to its stock of beliefs or remove one of the two aforementioned present beliefs from that stock. On Pollock's account, the proposed norm has the same status as a statement about the expected behaviour of any complex physical system. If true, it is exactly parallel to the following norm for the behaviour of the contemporary automobile: if an automobile is moving forward in a straight line and its driver turns its steering wheel to the left, the automobile should turn towards the left. This norm is just an empirical description of how the physical system generally functions, providing there are no unusual conditions such as a very slippery road surface or a break in the steering system. It is another question, on the design level, whether the direction of motion of automobiles should be controlled by a steering wheel operated by a manipulator who faces forward and sits off centre. A purely naturalistic approach to human rationality does not address such questions about the adequacy of the design, perhaps assuming that the operation of natural selection on the transmission of genetic information to contemporary human beings has ensured that human beings are optimally rational (in their competencies, if not in their performances). But evolutionary theory provides no such Panglossian reassurance, particularly if we pay attention to such human "spandrels" as the appendix or male nipples: who knows what similar excrescences lurk in human procedural knowledge of how to reason, perhaps making it in some respects irrational? Indeed, reliance on evolutionary theory to justify the correctness of human rationality would beg the question, since the development and justification of evolutionary theory has itself proceeded according to implicit norms of rationality which have been assumed to be correct. The argument against using evolutionary theory to ground norms of rationality recapitulates Frege's arguments against psychologism with respect to the foundations of logic.

Fortunately for Pollock, he has a second approach to constructing norms of rationality: a design approach (1995: 3). On this approach, rationality is a solution to certain design problems. The basic design problem is for the rational agent to interact with its surroundings so as to keep itself in existence and (in the case of humans) to perpetuate its species (1995: 6, n. 3); one might add that humans also manifest an impulse to understand how in general the world works, even when this understanding appears to contribute nothing to the survival of an individual or the perpetuation of the species. Pollock argues that much of the general architecture of rational thought, including human rational thought, can be inferred by taking a design stance. On specific details, however, it may turn out that human rational competencies are not optimal, or at least that there are defensible alternatives. Thus a less purely descriptive approach to rationality becomes possible.

2. Grades of cognition and the limits of rationality

On Pollock's account, there is a gradation among agents in how they act on

their environment to make their survival more likely (1995: 8-12). The simplest non-cognitive agent merely reacts to perceptual input; it does not learn. A more complex non-cognitive agent can acquire new reflexes through operant conditioning, which requires that the agent have built-in likes and dislikes for types of situations. *Epistemic cognition* forms mental representations of the world through *epistemic reasoning*. Simple cognitive agents would react to such representations through built-in reflexes, perhaps supplemented by conditioning. A more complex cognitive agent would have *practical cognition* directing its activity on the basis of both its beliefs about its situation and its likes and dislikes; the *practical reasoning* involved in such direction may vary in complexity from deciding on a single act to formulating a plan which takes a lifetime to execute.

As Pollock points out, it would be impossible for any agent to accomplish its design objective purely through epistemic and practical reasoning. Such reasoning is too slow; an agent which had to stop and think about what to do when it put its hand on something very hot would suffer a severe burn and probably permanent injury before the ratiocination was completed. Hence even a rational agent needs what Pollock calls *Q&I modules* (1995: 10), quick and inflexible routines for belief formation and practical cognition. Among human Q&I modules for belief formation are the computation of trajectories, much ordinary inductive inference, and intuitive integration. Practical Q&I modules include feature-liking; such built-in optative dispositions as desires to alleviate hunger, avoid pain, and pursue pleasure; a conditioning mechanism for forming new optative dispositions; and emotions like embarrassment and indignation.

Such Q&I modules have the advantage of speed but the disadvantage of inflexibility. Rationality is slow but flexible; it allows an agent to form beliefs and direct its actions when Q&I modules do not apply or when there is an explicit reason to override their output. A *fully epistemically rational agent* gives ultimate priority to rationality over Q&I modules in belief formation, while a *fully practically rational agent* gives such ultimate priority to rationality in the direction of its action. A *fully rational agent* is both fully epistemically rational and fully practically rational; Pollock thinks that humans are closer to full epistemic rationality than they are to full practical rationality.

3. Practical rationality

A standard model of practical rationality, which Pollock traces back to Hume (1995: 33) but which can already be found in Aristotle, supposes that practical reasoning arrives at decisions to act on the basis of a combination of beliefs and desires. Pollock argues convincingly (1995: 12-35) that this model is far too simple. He argues that practical reasoning requires seven distinct types of states: beliefs, situation-likings, feature-likings, intentions, and three kinds of desires (primitive, instrumental, present-tense action).

Situation-likings are fundamental. The function of rationality, Pollock supposes, is to make the world more to its possessor's liking. Hence a rational agent must have a way of telling how likable a situation is–a feeling produced by the agent's situation as the agent believes it to be. Humans are introspectively aware of such feelings.

Intentions encode the adoption of a plan. Planning involves constructing or discovering courses of action that might lead to the world's being more likable than otherwise. A rational agent will adopt a plan whose expected situation-liking is determined by deliberation to be at least as great as that of any of the competing plans under consideration. Ideally, a rational agent choosing among plans would consider each possible outcome of implementing each plan, estimate the probability of each such outcome given adoption of the plan, evaluate how likable that outcome would be, and adopt a plan whose weighted average of outcome likability was no lower than that of any other plan under consideration. A possible outcome is a type of situation characterized by certain features, whereas an agent's primitive likings and dislikings are for situation-tokens; the likability of a possible outcome is thus an expected likability, a weighted average of the likability of token situations of that type. To arrive at such an expected likability requires a cardinal measure of the likability of token situations. I have been somewhat surprised to discover that the humans I have consulted have no difficulty answering the question, "On a scale of 1 to 10, how good are you feeling right now?" Pollock however proposes to construct a cardinal measure indirectly, on the basis of a "quantitative feel" of a comparative preference relation among four arbitrarily chosen situations; he thinks that humans can introspectively tell whether they prefer situation B to situation A more than they prefer situation D to situation C. Further mathematical manipulation, combined with some assumptions about the preference relation, will produce from these data a cardinal measure allowing for unique comparisons of expected likabilities.

Feature-likings are a shortcut required by constraints of time and resources. Theoretically, a rational agent could work out by reasoning what features of situations are causally relevant to their being liked or disliked. In practice, the agent has to act before having time to go through the elaborate reasoning that would be required. Hence a rational agent needs Q&I modules which provide this information. Pollock speculates (1995: 20) that humans acquire feature-likings through their ability to imagine situations (which must be types rather than tokens) and respond conatively to them; equally speculatively, we can conjecture that humans recognize directly in a token situation those aspects of it which they like or dislike–but perhaps what appears to be immediate recognition is a product of learning. Parenthetically, Pollock notes that there could be a rational agent for whom feature-likings are fundamental; such a rational agent would need, Pollock argues, both a cardinal measure of primitive feature-likings and a way of computing a liking for combinations of features from the likings of individual features (1995: 20-21). Humans seem to use Q&I modules to compute the comparative expected value of plans on the basis of situation-likings and feature-likings; Pollock thinks that artificial

rational agents might be able to solve the integration problems required for this computation explicitly.

Primitive desires encode goals and initiate planning. Goals, construed as combinations of features, are required for planning by limitations of time and resources. Starting with a specific goal is necessary for efficient interest-driven epistemic reasoning, as opposed to a time-consuming random generation and evaluation of plans. A plan which can attain a goal can be presumed to have a positive expected value if the expected likability of the goal's combination of features is greater than the expected likability of the situation that would otherwise result. But this presumption can be defeated by other features of the situation that results from carrying out the plan. Considerations of feasibility require that a rational agent not only form desires as a result of epistemic reasoning about the expected likability of certain combinations of features, but also have Q&I modules which propose goals and produce their default adoption, unless the agent's reasoning judges them unsuitable. Humans have such *optative dispositions* to try to alleviate hunger, avoid pain and pursue pleasure. Conditioning can lead to new optative dispositions. In a fully practically rational agent, reasoning that a desired goal is unsuitable would extinguish the desire, and reasoning that a goal is suitable would produce a desire for it; Pollock notes drily (1995: 27-28) that humans are not fully rational in either of these respects.

Instrumental desires are produced by adoption of a partial plan (for example, getting a copy of this paper to my commentator's philosophy department by Tuesday morning as a way of achieving the goal of his receiving it before he leaves on Wednesday for a conference); such desires initiate further planning.

Present-tense action desires are needed to initiate action, since adopted plans may leave the scheduling of steps indefinite. Action-initiating desires may be produced by optative dispositions or by the adoption of a plan. When present-tense action desires conflict, an agent will act on the strongest of these desires. Thus a rational agent will proportion the strength of such a desire derived from an adopted plan to the expected likability of the tail of the plan, that part of it which remains to be carried out. Pollock seems to assume that the strength of desires produced by optative dispositions (e.g. a human being's disposition to try to alleviate its hunger) will also be proportional to the expected value of satisfying them, because he thinks a rational agent should at any given time perform the action it most wants to perform (1995: 31). But this assumption seems implausible; a human being may for example have a fierce desire to drink or eat what is in front of him or her and a weak desire to postpone the satisfaction of this desire (for example in an extreme situation where survival requires rationing a limited supply). There seems to be a need in a fully practically rational agent to override a strong present-tense action desire due to an optative disposition in the light of a rationally based judgement that some alternative action has greater expected value; Pollock (1995: 35) seems to assume that such reasoning would dispel the suboptimal desire in a fully rational agent, but overriding it would also seem to be rational.

Pollock identifies four important consequences of the above-described conception of practical rationality:

(1) As previously mentioned, practical reasoning requires more states than beliefs and desires. It needs also situation-likings, feature-likings and intentions. And desires are of three types--primitive, instrumental and present-tense action.

(2) A real agent cannot rely on ratiocination for all its practical cognition. Constraints of time and resources require it to use Q&I modules, whose output is correctable by ratiocination in a fully rational agent.

(3) All types of evaluative attitudes other than situation-likings are subject to rational criticism. Feature-likings are irrational if the expected likability of a situation possessing the feature(s) is relatively low. Instrumental desires can be criticized by evaluating the plan from which they are derived. Primitive desires, whether produced by optative dispositions or by ratiocination, can be criticized on the ground that the goal they encode does not have a high relative expected value. Present-tense action desires can be criticized (if they arise from adoption of a plan) by evaluating the plan from which they are derived or (if they arise from an optative disposition) by arguing that fulfilling them does not contribute to living a good life, in the sense of a life in which the agent's situation-tokens are more likable than otherwise.

(4) Contrary to Hume, not all reasoning is epistemic. Pollock's model includes three types of non-epistemic state transitions which are subject to rational evaluation: (a) from beliefs about the expected situation-likings of potential goals to desires (adoption of goals), (b) from beliefs about the relative values of plans to intentions (adoption of plans), and (c) from choosing the strongest present-tense action desires to actions.

4. The interdependence of epistemic and practical cognition

Pollock argues that epistemic and practical cognition are distinct types, but that they cannot be discussed in isolation because of the complex interdependence between them. Attempts to reduce epistemic cognition to practical cognition, by identifying it with practical cognition about what to believe, founder on an infinite regress generated by the fact that all practical cognition depends on beliefs about the agent's current situation. Attempts to treat epistemic cognition in isolation from practical cognition founder on the fact that all epistemic cognition is interest-driven. "The whole point of epistemic cognition is to help the agent solve practical problems." (195: 36) This claim strikes me as overly restrictive, since some human beings sometimes deploy epistemic reasoning simply for the sake of acquiring a general understanding of the world, with not even an indirect connection to practical interests. Indeed, such a theoretical orientation is responsible for the origins of western science in the ancient Greek cosmologists, who were simply curious about how the

cosmos came to be, what shape the earth has, where it is in the cosmos, why it stays where it is, and so on–with no ulterior practical interest in mind. Nevertheless, Pollock is correct that the epistemic reasoning of a rational agent is not just random derivation of conclusions from premisses, but is driven by interests, whether practical or theoretical. He refers (1995: 36) to the questions posed by practical cognition as *ultimate epistemic interests*. Not only does practical cognition direct epistemic cognition, but epistemic reasoning provides goals for practical cognition, encoded as *epistemic desires*, e.g. the desire to know what time it is, which in general is satisfied not by epistemic reasoning but by forming a plan and carrying it out. This interdependence of epistemic and practical cognition is the central feature of the architecture for rational cognition implemented in OSCAR.

5. Reasoning, inference graphs, and arguments

Epistemic reasoning, on Pollock's account, proceeds from perceptual input or previously held beliefs to further beliefs. Pollock understands by an *argument* the record of the transitions involved in a sequence of reasoning; Pollock's "argument" is thus what is sometimes called complex argumentation, but without any communication to an audience or with an interlocutor. A single transition in such a sequence takes place from one or more reasons to a conclusion in accordance with what he calls a *reason schema*, for example: 'x looks red to me' is a reason for me to believe 'x is red'; Pollock's "reason schemas" correspond to what are called "argument schemes" or "argumentation schemas" in the argumentation literature. A *linear argument* can be viewed as a finite sequence of propositions, each a member of *input* (the premises not inferred from any reason) or inferable from previous members of the sequence in accordance with a reason schema. Pollock insists, against some models of argumentation structure in the artificial intelligence literature, that there are non-linear arguments, in particular arguments that discharge suppositions through conditionalization, reductio ad absurdum or reasoning by cases (1995: 88). Whereas the structure of an argument records the actual sequence of inferences by a reasoning agent, Pollock uses *inference graphs* to record relations of dependence, which can be compatible with more than one sequence of inferences; for example, deducing from a conjunction each of its conjuncts, then reconjoining them, gives rise to a single inference graph, regardless of the order in which the conjuncts are deduced (1995: 87). Pollock has a perspicuous solution to the problem of how to represent the distinction between linked and convergent reasoning, or in another terminology between coordinatively compound and multiple argumentation. A single inference to a conclusion from more than one reason is represented by an arrow coming from each reason to a single node for the conclusion; this diagrams linked reasoning (also called coordinatively compound argumentation), where the reasons work together to support the conclusion. Two or more independent arguments for a conclusion are represented by having a separate node for each supporting separate

argument; this diagrams convergent reasoning (multiple argumentation). This graphical technique allows Pollock to associate unique strengths with nodes and to regard different nodes as defeated by different defeaters (1995: 88-89).

Pollock argues that the reasons used in epistemic reasoning are neither linguistic items nor propositions but mental states, including perceptual images as well as beliefs (1995: 54-55). His reason is that human beings form beliefs directly on the basis of perceptual input, without having any beliefs about that input; in fact, humans rarely have beliefs about what the things they see look like, what the things they hear sound like, etc. In cases where the reasons are beliefs, we can say, loosely, that 'P & Q' is a reason for 'P', when what we mean is that believing 'P & Q' is a reason for believing 'P'.

One of the difficult problems for representing argument structure in a standard format is how to accommodate suppositional reasoning. Pollock writes arguments in a vertical sequence, numbering each proposition sequentially, and writing before a supposed reason the word "suppose", and before an inferred conclusion "from" followed by the number of each reason used in inferring it. Here is an example:

1. Suppose prob(*F/G*) $ *r* & *Gc*.

2. Suppose (*p* / -*Fc*).

3. *Fc* from 1.

4. -*p* from 2,3.

This is essentially the system used in contemporary formal logic texts for exhibiting natural deduction proofs, but without any of the devices used in such texts for tracking the dependency of inferred conclusions on particular suppositions. In his system of inference graphs, however, Pollock tracks the dependency by building a supposition into each node at which the supposition is used (1995: 88). Thus a node encodes an inference to a *sequent*, an ordered pair <*X*, *p*> whose first member *X* is a (possibly empty) set of propositions and whose second member *p* is a single proposition inferred relative to *X* as supposition. The inference graph corresponding to the above argument would have as its first node the ordered pair <{prob(*F/G*) $ *r* & *Gc*}, *Fc*>, and as its second node, with an arrow descending to it from the first node, the ordered pair <{prob(*F/G*) $ *r* & *Gc*, (*p* / -*Fc*)}, -*p*>.

6. Defeasible reasoning

Pollock sets himself firmly against deductivism with respect to epistemic reasoning. That is, he opposes the view that the only good reasons for drawing a conclusion are conclusive reasons, reasons which logically entail the conclusion. In this respect, he agrees with the current trend in the artificial

intelligence literature, which has many systems of so-called non-monotonic reasoning, reasoning in which new information can make it rational to cease to draw the conclusion, even though the inference was perfectly good at the time and the reasons on which it was based remain unchallenged. The artificial intelligence literature, however, seems to regard such non-monotonic reasoning as an unfortunate consequence of avoidable incompleteness of information in a database. Pollock argues on the contrary that what he calls *defeasible reasoning* is normal and quite unavoidable: "It is logically impossible to reason successfully about the world around us using only deductive reasoning." (1995: 41) Reasoning from the way things appear to the way they are is unavoidably defeasible; it is impossible to construe physical object statements as logical constructions from phenomenalist claims. Reasoning from observed regularities to a universal generalization is unavoidably defeasible. Reasoning from 'most *A*s are *B*s and this is an *A*' to 'this is a *B*' is unavoidably defeasible. "Almost everything we believe is believed at least indirectly on the basis of defeasible reasoning, and things could not have been any other way." (1995:42) To me these remarks are an absolutely convincing refutation of deductivism.

"Defeasible reasoning" is so-called because it can be defeated. Even though the conclusion follows by a legitimate reason schema and the reasons remain unchallenged, new information may defeat the reasoning. An important contribution of Pollock's, almost universally adopted in the artificial intelligence literature but hardly noticed in the informal logic and argumentation literature, is that there are two kinds of defeaters. Where *P* is a *prima facie reason* (as opposed to a conclusive reason) for *Q*, *R* is a *rebutting defeater* for *P* iff *R* is a reason for denying *Q*. This is the type of defeater with which we are familiar. For example, that I am keeping a sword for my friend is a prima facie reason for giving it back to him when he asks for it, but that he intends to do harm to himself with the sword is a reason for not giving it back to him when he asks for it. My friend's intention to harm himself defeats the reason that I am keeping the sword for him. It does so by being a reason for the contradictory conclusion. The other, less familiar kind of defeater Pollock identifies is what he calls an *undercutting defeater*. Undercutting defeaters attack the connection between the prima facie reason and the defeasibly drawn conclusion rather than attacking the conclusion directly; where *P* is a *prima facie reason* (as opposed to a conclusive reason) for *Q*, *R* is an *undercutting defeater* for *P* iff *R* is a reason for denying that *P* would not be the case unless *Q* were the case. For example, that an object looks red to me is a prima facie reason for believing that it is red. The information that the object which looks red to me is illuminated by red light undercuts the prima facie reason, because it makes it unreasonable to infer from the object's looking red that it is red; when illuminated by red light, objects of many colours look red. But this defeater does not give me any reason to believe that the object is not red; it is not a rebutting defeater.

7. Strength of justification

How strongly does a sequence of reasoning support a conclusion to which it leads? Pollock's answer to this question is surprising, but I can find no fault with the reasoning by which he gets to it. He begins by noting, sensibly enough, that we need to assign strengths to the *input* states on which such reasoning is based as well as to the conclusive or prima facie reasons used at each inference link. A perceptual input can be more or less strong; for example, an object can look more or less clearly red (1995: 101). To assign degrees of strengths to reasons, Pollock uses as a standard of comparison instances of the statistical syllogism: If $r > 0.5$ then 'prob(*F/G*) $ *r* & *Gc*' is a prima facie reason for '*Fc*', the strength of the reason being a monotonic increasing function of *r* (1995: 93). Rather than taking *r* itself to be the measure of the strength of the reason, Pollock takes $2 \times [r - 0.5]$, which has the consequence that the strength of an instance of statistical syllogism is 0 when $r = 0.5$ and 1 when $r = 1$; thus the strengths of reasons range between 0 and 1. The strength of prima facie reasons other than instances of the statistical syllogism is measured by taking those instances as a standard of comparison, in the following way: "If *X* is a prima facie reason for *p*, the strength of this reason is $2 \times [r - 0.5]$, where *r* is a real number such that an argument for ¬*p* based upon the suppositions 'prob(*F/G*) $ *r* & *Gc*' and '(*p* / -*Fc*)' and employing the statistical syllogism exactly counteracts the argument for *p* based upon the supposition *X*." (1995: 94) The consequence that all reasons can be linearly ordered by strength is natural, but conflicts with some recent proposals on nonmonotonic reasoning.

If an argument combines several inferences using reasons of less than unit strength, how is the degree of support for the ultimate conclusion to be computed? The usual answer is to treat the degree of strength of each reason as a probability, and to use the probability calculus to compute how strongly the *input* supports the ultimate conclusion. The ultimate conclusion is regarded as justified only if it is made sufficiently probable by the cumulative reasoning. Pollock offers a powerful argument against this *generic Bayesianism*. He notes that, for the Bayesian, inference rules can be applied blindly, without making probability calculations, only if it follows from the probability calculus that the probability of the conclusion licensed by the inference rules is greater than or equal to the minimum of the probabilities of each premiss to which it is applied; such an inference rule he describes as *probabilistically valid* (1995: 95-96). But in general deductively valid rules of inference from multiple premisses are not probabilistically valid; if more than one premiss has a probability less than 1, then according to the probability calculus the conclusion will have a probability less than that of any premiss. This means that, if an engineer combines 100 pieces of information, each with a probability of .99, in a chain of deductive reasoning used to compute the size of a girder for a bridge, her conclusion would have a probability too small to be justified–which means that it would be impossible to build bridges. Even more damaging than this counter-intuitive result is the consequence that generic Bayesianism is

self-defeating: a Bayesian reasoner cannot compute the probabilities required to decide whether to hold a belief. Suppose, for example, that a Bayesian reasoner holds the following beliefs:

prob(*P*w*Q*) = prob(*P*) + prob(*Q*) ! prob(*P*&*Q*)

prob(*P*) = 0.5

prob(*Q*) = 0.49

prob(*P*&*Q*) = 0.

In the typical cases to which belief updating is relevant, the second and third premisses will have a probability less than 1. Hence the inference to the conclusion that prob(*P*w*Q*) = 0.99 is not probabilistically valid. The Bayesian reasoner has no way to calculate the probability of the conclusion. Trying to replace the second and third premisses by their conjunction leads to an infinite regress, since the same difficulty recurs in trying to compute the probability of the conjunction using conditional probability (1995: 96-98). Thus generic Bayesianism is according to Pollock incoherent. Its mistake is to treat epistemic "probability" as a probability in the sense of the probability calculus. Since deductive reasoning from multiple premisses preserves high epistemic "probability", this probability is not a probability in the sense of the classical probability calculus. Curiously, Pollock does not reply to the standard "Dutch book" argument that rational degrees of confidence in propositions must obey the laws of the probability calculus, or to the more sophisticated versions of that argument due to Ramsey and his followers. If Pollock is right, there must be a hitherto undetected flaw in these arguments.

Pollock's substitute for Bayesianism is what he calls the *weakest link principle*. Applied to deductive reasoning, this is the principle that the degree of support of the conclusion is the minimum of the degrees of support of a deductive argument's premisses (1995: 99). Any account which makes the degree of support weaker than that minimum, Pollock notes, will be self-defeating in the same way as generic Bayesianism. And the weakest link principle seems to be the only principle that is not completely ad hoc which explains how adding inferences from new premisses weakens an argument. Pollock extends the weakest link principle to defeasible arguments by noting that, whenever *P* is a prima facie reason for *Q*, we can use conditionalization to construct an argument for (*P* e *Q*) with no premisses, whose conclusion is therefore supported to the same degree as the strength of the prima facie reason. This technique allows us to reformulate any defeasible argument so that its conclusion is a deductive consequence of *input* along with a number of conditionals so justified; hence, by the weakest link principle for deductive arguments, the degree of support for the conclusion is the minimum of the degree of justification of the members of *input* used in the argument and the strengths of the prima facie reasons (1995: 100-101). Thus the weakest link principle for defeasible arguments is that the degree of support of the

conclusion of a defeasible argument is the minimum of the strengths of the prima facie reasons used in it and the strengths of the *input* states to which it appeals (1995: 101).

Another surprising claim about strength of support is that two independent reasons for a conclusion do not provide stronger support than one (1995: 101-102). "If we have two separate undefeated arguments for a conclusion, the degree of justification for the conclusion is simply the maximum of the strengths of the two arguments." (102) Pollock does not so much argue for this principle as explain away apparent counter-examples. If the testimony of one person is confirmed by the independent testimony of another person, we generally take their joint testimony to support their claim more strongly than the testimony of either one without the other. But this increase in support is not due to there being two independent reasons for the same conclusion, Pollock claims. Rather, it is due to our having a single combined reason which is usually a stronger prima facie reason than the initial reason; in general, prob(*p* is true / S_1 asserts *p* and S_2 asserts *p* and S_1 … S_2) > prob(*p* is true / S_1 asserts *p*). In general, but not always. In a community where speakers tend to confirm each other's statements only where they are fabrications, the reverse is true. Anyone who thinks that accrual of reasons provides stronger support than the strongest of the reasons so accrued needs to consider whether apparent instances of such accrual are in fact of the form Pollock identifies. My own brief consideration of apparent counterexamples to his position suggests that he is right.

8. Justification and defeat

So far we have been considering the strength of support for a belief without attending to the possibility that a defeasible inference can be defeated. Pollock's position on when inferences are defeated is the most subtle of which I am aware. It has been worked out over several published iterations, each one accommodating cases which the previous one failed to solve satisfactorily. Pollock has used the most recent version of his theory (1994, 1995: 102-140) to produce congenial solutions to the lottery paradox and the paradox of the preface.

The account begins with an analysis of when one node in an inference graph rebuts or undercuts another. A rebutting node must have as its conclusion the negation [actually the contradictory–DH] of the node it rebuts. But in addition it must depend on suppositions which are a subset of those on which the node it rebuts depends. Otherwise one could rebut any defeasible argument simply by noting that the contradictory opposite of the argument's conclusion can be deduced from itself as supposition. Further, the strength of support for the conclusion of the rebutting node must be at least as great as the strength of

support for the node being rebutted. A weaker argument for an opposite conclusion, Pollock argues, does not diminish the strength of support from the original argument. Otherwise two such weaker arguments would suffice to rebut the original argument, and that is just the principle of accrual already rejected, that two independent arguments can have a strength greater than that of either of them. Cases where a weaker argument for an opposing conclusion seems to diminish the strength of support from the original argument are handled analogously to cases of apparent accrual of support from two independent arguments; we do not have two arguments, but one new one, with a different set of premisses. Putting all these points together leads Pollock to formulate the following definition of when one node " in an inference graph rebuts another node $: "(1) $ is a pf-node [node representing a conclusion reached by a prima facie reason–DH] of some strength > supporting some proposition *q* relative to a supposition K; (2) " is a node of strength 0 and supports ¬*q* relative to a supposition O, where O f K; and (3) 0 $ >." (1995: 103) The definition of when one node undercuts another node is analogous, except that an undercutting node supports ((p_1 & ... & p_k q *q*), where p_1, ..., p_k are the propositions supported by the immediate ancestors of $ (i.e. the conclusions of the sequents at the nodes from which $ is immediately inferred) and '((p_1 & ... & p_k q *q*)' means that it is not the case that (p_1 & ... & p_k) wouldn't be true unless *q* were true (1995: 86). Pollock encodes defeat relations between nodes in an inference graph in the form of defeat links, with a specific type of arrow going from the rebutting or undercutting node to the node it defeats.

The central concept in Pollock's account of justification and defeat is the assignment of a defeat status to a node in an inference graph, either "defeated" or "undefeated". The constraints on a status assignment are such that a given inference graph may be subject to different status assignments. A node in an inference graph is undefeated iff every status assignment assigns "undefeated" to it, defeated outright iff no status assignment assigns "undefeated" to it, and otherwise defeated provisionally. Because some inference graphs make it impossible to assign defeat statuses to all nodes consistent with the basic conception of a status assignment, Pollock admits partial status assignments which assign no status to some nodes (1995: 122-124). A status assignment is then a maximal partial status assignment, i.e. a partial status assignment which is not properly contained in another partial status assignment (1995: 124). A partial status assignment is an assignment of the status "defeated" or "undefeated" to a subset of the nodes of an inference graph which meets the following three conditions: (1) It assigns the status "undefeated" to every *d-initial node*, i.e. every node such that neither it nor any of its ancestors is the terminus of a defeat link. Second, it assigns "undefeated" to every node to whose immediate ancestors it assigns "undefeated" and to whose defeating nodes it assigns "defeated". Third, it assigns "defeated" to every node which has an immediate ancestor to which it assigns "defeated" or a defeating node to which it assigns "undefeated". (1995: 123-124)

A conclusion is justified if and only if it is supported by an undefeated node of the inference graph of a sequence of epistemic reasoning. It is justified to any degree up to and including the strength of the strongest undefeated node which supports it.

9. Justification and warrant

Whether a belief is justified, and to what degree, is a function of how far a reasoner who entertains the belief has got in its reasoning. Even without new input, further reasoning can produce a conclusion which rebuts or undercuts a previously undefeated node; hence, what was justified becomes unjustified. Conversely, an undefeated defeater can become defeated, so that what was unjustified becomes justified.

Pollock uses the term *warrant* to talk about whether a sequent is ultimately justified by an indefinitely continuing sequence of reasoning. A sequent is warranted to a certain degree by a certain *input* iff it is justified to that degree at every stage in the sequence of reasoning from *input* after some stage (1995: 133). It is *ideally warranted* to a certain degree by a certain *input* iff the set of all nodes producible by the reasoner includes an undefeated node at least that strong which supports the sequent. Curiously enough, warrant and ideal warrant are not equivalent (1995: 133-134).

10. Conclusion

I have touched on only some aspects of Pollock's theory of rationality. Other aspects relevant to informal logic, critical thinking and the theory of argumentation include reflexive cognition (1995: 43-46), acceptance rules, the statistical syllogism and its generalizations, definite and indefinite probabilities, direct inference, enumerative induction, statistical induction, the control structure for interest-driven reasoning (monotonic and defeasible) which reasons both forwards from the *input* and backwards from the desired conclusion, and plan-based practical reasoning (1995: 140-298). In the end, Pollock is able to summarize his architecture for rational cognition in a single diagram (1995: 304), which of course requires extensive interpretation.

Pollock has often changed his position on the nature of rationality, in response to difficulties that he found in previous accounts. In fact, one of the virtues of computational implementation, in this area as in others, is that running a program discloses problems that might not (indeed, did not) strike a critical human reasoner. Pollock's work has the complementary virtue, often missing in "seat-of-the-pants" artificial intelligence research, of philosophical sophistication about the foundations of the computational implementation.

Pollock's plan in his OSCAR project is to produce a general architecture for

rationality and an inference engine for interest-driven defeasible reasoning (which he has already done), then use it to analyse such specific kinds of reasoning as abduction (inference to the best explanation), reasoning by analogy, causal reasoning, temporal reasoning, spatial reasoning, reasoning about change and persistence, and reasoning about other rational agents. He is distributing OSCAR in order to encourage investigators to use it in their investigations and provide feedback on its design (1999: I-5). I believe that those of us with interests in norms and control structures for specific types of reasoning should take him up on his offer.

References

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

Pollock, John L. (1987). "How to build a person: the physical basis for mentality." *Philosophical Perspectives* 1, *Metaphysics*: 109-154.

Pollock, John L. (1988). "The building of Oscar." *Philosophical Perspectives* 2, *Epistemology*: 315-344.

Pollock, John L. (1989). *How to Build a Person: a Prolegomenon*. Cambridge, MA: MIT Press.

Pollock, John L. (1990a. *Nomic Probability and the Foundations of Induction*. New York and Oxford: Oxford University Press.

Pollock, John L. (1990b). "A theory of defeasible reasoning." *International Journal of Intelligent Systems* 6: 33-54.

Pollock, John L. (1991). "Self-defeating arguments." *Minds and Machines* 1: 367-392

Pollock, John L. (1992). "How to reason defeasibly." *Artificial Intelligence* 65: 1-42.

Pollock, John L. (1994). "Justification and defeat." *Artificial Intelligence* 67: 377-408.

Pollock, John L. (1995). *Cognitive Carpentry: A Blueprint for How to Build a Person*. Cambridge, MA: MIT Press.

Pollock, John L. (1997). "Reasoning about change and persistence: a solution to the frame problem." *Nous* 31: 143-169.

Pollock, John L. (1998). "Procedural epistemology." In Terrell Ward Bynum and James H. Moor (Eds.), *The Digital Phoenix: How computers are changing*

philosophy. Oxford: Blackwell.

Pollock, John L. (1999). *Download OSCAR*. http://www.u.arizona.edu/~pollock/oscar.html. Visited on April 24.