Pollock on Practical Reasoning

David Hitchcock *McMaster University*

Abstract: The epistemologist John Pollock has implemented computationally an architecture for a rational agent which he calls OSCAR. OSCAR models both practical and theoretical (or epistemic) reasoning. I argue that Pollock's model of practical reasoning, which has seven components, is superior not only to the two-component belief-desire model stemming from Aristotle, but also to the three-component belief-desire-intention model developed especially by Michael Bratman. Despite its advantages, Pollock's model of practical reasoning is incomplete in at least three respects: it is solipsistic, it is egoistic and it is unsocial.

Résumé: L'épistémologue John Pollock a mis en exécution une structure de rationalité qui représente les raisonnements pratique et théorique (épistémique). J'avance que ce modèle de la raison pratique, formé de sept parties composantes, est supérieur non seulement au modèle provenant d'Aristote, fondé sur les croyances et des désires, mais aussi au modèle basé sur les croyances, les désires et les intentions, développé principalement par Michael Bratman. Malgré ces avantages, le modèle de la raison pratique de Pollock est incomplet pour au moins trois raisons: il est solipsistique, égotiste, et non sociale.

Keywords: reasoning, practical reasoning, John L. Pollock, desires, beliefs, intentions, likings, Hume, Michael E. Bratman

1. Introduction: the nature of practical reasoning

By "practical reasoning" I shall understand reasoning about what to do. Practical reasoning is to be contrasted with reasoning about what to believe, which is often called "theoretical reasoning", but which I shall here call "epistemic reasoning", following Pollock (1995: 9). Doing something includes as the most elementary case (1) a simple physical action, such as raising one's arm. More complex cases are (2) a series of physical actions and (3) adoption of an intention to perform some action later (a "plan", which may be only partially elaborated at first). Typically the actions to be performed will not be described in terms of how the agent moves the parts of its body (and indeed it may be somewhat indeterminate what bodily movements will constitute the action), but rather in terms of what function the movement of the parts of the body will amount to, e.g., calling someone on the telephone and asking them a certain question. (4) Plans may be logically complex, including for example disjunctions or conditions. (5) More general than a plan is a policy, which is the carrying out of a certain type of plan whenever specified

© *Informal Logic* Vol. 22, No. 3 (2002): pp. 247-256.

conditions obtain (e.g., whenever I want to walk across a street, do so only when it is safe). More complex still are (6) cases where the agent is not an individual human being but an organization of human beings—an academic department, a municipal government, the board of directors of a joint-stock company, the executive of a voluntary organization, etc. Actions of all these types include intentional omissions, i.e., intentionally not initiating (now or later) a certain bodily movement or series of bodily movements, defeating a resolution to undertake some initiative, etc. Thus, generically, practical reasoning is reasoning directed to the adoption of a policy by some agent, where policies include as limiting cases plans (policies adopted for just one occasion) and actions (plans with only one component), and the agent adopting the policy may or may not be identical to the set of individuals carrying out the reasoning.

On the face of it, the criteria for good practical reasoning must be different from those for epistemic reasoning. For we evaluate the inferential link in epistemic reasoning by considering how likely it is that the conclusion is true if the premises are true: the inference is deductively valid if it is necessary that the conclusion is true if the premises are true, inductively strong if it is probable that the conclusion is true if the premises are true, and provisionally valid if *ceteris paribus* the conclusion is true if the premises are true. But the conclusion of practical reasoning is a policy decision, which is not the sort of thing that can have a truth-value. Policy decisions can be wise or foolish, far-sighted or short-sighted, and so on, but they cannot be true or false. There is no such thing as a true policy or a false policy. One can attempt to assimilate the imperative conclusions of practical reasoning ("Let's invite them to dinner next Saturday") to the indicative conclusions of epistemic reasoning by recasting those imperative conclusions as indicative "ought" statements ("We ought to invite them to dinner next Saturday"). But this assimilation will not work, for two reasons. First, it is doubtful that "ought" statements have truth-values. If we adopt a reistic conception of truth according to which any true assertive is true in virtue of some truth-maker, such as a fact or an event or a state of affairs, then we are faced with the problem of finding a truth-maker for supposedly true "ought" statements. In virtue of what state of affairs could it be true that we ought to invite some friends for dinner next Saturday? Empirical investigation can discover facts relevant to our decision-making—for example, that we have not seen these friends for some months and that our calendar is empty for that day. But it cannot discover that we ought to invite them for dinner. If there is a fact that we ought to invite our friends for dinner, it is a queer sort of fact indeed. More likely, there is no such fact, and it is not true that we ought to invite our friends for dinner. Which is not to say that there is the opposite fact that it is not the case that we ought to invite our friends for dinner; there is no such opposite fact, and it is not true that it is not the case that we ought to invite our friends for dinner. Second, and more decisively, it is always possible without contradiction to affirm an "ought" statement and make the opposite policy deci-

sion. We can consistently say, for example, "We ought to invite our friends to dinner next Saturday, but let's not." "Ought" statements are not the same as policy decisions.

2. The belief-desire model of practical reasoning

The simplest model of good practical reasoning is the belief-desire model first articulated by Aristotle. According to the model prescribed in his *Nicomachean Ethics* (III.3.1112b15-20), good deliberation begins with a wish for some end. The practical reasoner then considers how this end is to be attained. Having found a means of attaining it, the reasoner then considers how this intermediate end is to be attained, and so on until some means is discovered which is an action within the person's power. The conclusion of the reasoning is a decision to perform this action, which in Aristotle's model immediately issues in the action itself. If at any stage the agent discovers more than one means of achieving a desired end, the agent looks for the easiest and finest of them; thus Aristotle incorporates considerations of efficiency and nobility in his model. For other descriptions by Aristotle of practical reasoning as a process of reasoning from a desire for some end via beliefs about the means of achieving it, see *On the Soul* III.10.433a13-20, *Nicomachean Ethics* VI.2.1139a32-b5, and *Eudemian Ethics* II.10.1227a2-30. A variant formulation holds that practical reasoning combines a universal judgement about what ought to be done (e.g., that everything sweet ought to be tasted) with a particular judgement or judgements bringing the present situation of the agent under the universal judgement (e.g., that this is sweet) (see for example *On the Soul* III.11.434a17-19, *Movement of Animals* 7.701a7-24, and *Nicomachean Ethics* VII.3.1147a25-31). This variant can be assimilated to the means-end model by allowing an end to be achieved in the very performance of an action; for example, eating a particular piece of chocolate could be construed as a means of attaining the generic end of tasting something sweet. The universal "ought" judgements in the variant model must be construed as expressions of a desire, in order to make the variant model consistent with Aristotle's repeated claim that practical reasoning requires a desire for some end to initiate it.

The classic modern statement of the necessity of desire for practical reasoning occurs in David Hume's *Treatise of Human Nature* II.3.3. In Hume's famous words, "Reason is, and ought only to be the slave of the passions." Passions always involve desires, which are the initiator of practical reasoning; neither abstract reasoning nor inductive reasoning about causes and effects can by itself initiate action. Passions are unreasonable only when they are founded on a false supposition or choose a means insufficient for the desired end. Nothing enters into practical reasoning but a desire for some end and beliefs about the means of achieving it.

3. The belief-desire-intention (BDI) model of practical reasoning

Among contemporary philosophers, Michael Bratman (1987) has made a new contribution to the traditional model of practical reasoning by arguing convincingly for a third component: intentions, which are paradigmatically future-directed. The traditional model explained intentions adverbially: to do something with an intention to do it is to do it intentionally, where doing something intentionally is construed as doing something in terms of the agent's desires and beliefs. Intentions to do something in the future were reduced to appropriate desires and beliefs. Bratman resists this reduction. To form an intention to do something in the future, he argues, is to adopt a plan, which typically is partial. Human beings, unlike many nonhuman animals, are planning agents; they need to adopt plans for the future in order to allow their reasoning about what to do to reach beyond the present moment and to coordinate their activities with each other and with those of other people. We do not do justice to this important aspect of human rationality if we treat future-directed intentions as a mere construct of present desires and beliefs. Rather, we need a planning theory of intention which articulates the regularities and norms in virtue of which intentions are a mental attitude distinct from desires and beliefs. Intentions are an output of practical reasoning, and also an input to future practical reasoning, in the form for example of a constraint on admissible options. In later work, Bratman (1999) has fleshed out his original theory, for example in a discussion of when it is rational to reconsider previously adopted plans.

The belief-desire-intention (BDI) model of practical reasoning has been implemented in a number of computer-based decision support systems.

4. Pollock's belief-desire-intention-liking (BDIL) model of practical reasoning

Pollock (1995) incorporates Bratman's intentions, but adds a new type of component to practical reasoning, which he calls "likings". Furthermore, he distinguishes two types of likings and three types of desires, which in combination with beliefs and intentions produce a seven-component model of practical reasoning.

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. Pollock 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 (and to accumulate the experience needed as inputs to such reasoning). Hence a rational agent needs Q&I (quick and inflexible) 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 pre-

sumption can be defeated by other features of the situation that result 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 this paper to the editor of this issue by the promised deadline as a way of achieving the goal of his including it in the issue); 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.

5. Strengths of Pollock's BDIL model

A great strength of Pollock's model is its recognition that desires are not the ultimate canon of appeal in practical reasoning. Contrary to Hume, a desire can be subject to rational criticism, on the ground that satisfaction of the desire will produce a situation less to the agent's liking than some alternative option. This point is a matter of common sense once it is articulated; it is implicit, for example, in the common recognition that people in the grip of a harmful addiction would be better

off if they did not have the desire for the addictive experience. Addicts often recognize this fact themselves. Philosophical theories of practical reasoning, perhaps under the influence of Hume, have tended not to allow for it. They have recognized that pleasant and painful experiences cause desires and aversions (see for example Aristotle's *On the Soul* III.7.431a8-10 and Hume's *Treatise* II.3.3 (1975: 414)). But Hume in particular left no room for the rational assessment of desires according to the pleasure to be gained from satisfying them. (Aristotle does have a theory of correct and incorrect desires, but exploration of his theory would take us too far afield.)

A further strength of Pollock's theory is his use of the degree to which a token situation is likable as the ultimate touchstone of practical reasoning, rather than appealing to how pleasant or painful the situation is to the agent. Pollock's formulation is better in two respects. First, the concepts of pleasure and pain are too easily construed simply in terms of gratification of the appetites connected with the senses of touch and taste. Such comforts and delights are certainly some part of living a good life, but they are not the whole of it. As John Stuart Mill memorably put it, "I had rather be Socrates dissatisfied than a pig satisfied" (Mill 1888). A Pollockian rational agent would express the point in terms of likings rather than preferences: I would like it more if I were Socrates dissatisfied than if I were a satisfied pig. Pollock's theory, unlike Mill's, does not prescribe any particular hierarchy of situations. But, in taking personal situation-likings as basic, it allows each agent to accommodate the preference expressed by Mill. Second, Pollock's theory appeals not to how much an agent actually likes a token situation but to how much the agent would like the situation if the agent's relevant beliefs were correct. Thus token situation-likings become subject to rational criticism in terms of the correctness of the beliefs which produce them. Recognition of this sort of rational criticism in a theory of practical reasoning is not new; even Hume acknowledged it, in his case with reference to "passions", i.e., desires. But it is less common to allow it in a theory which takes as basic some analogue to Pollock's situation-likings.

Another strength of Pollock's model is his recognition that a rational agent operating in real time in a hazardous environment with quite limited computational resources needs quick and inflexible (Q&I) modules to generate actions by default in many situations. Without the reflex reaction of withdrawing one's hand immediately from painful contact with a flame or similarly hot object, human beings would find the world much less to their liking than they now do. Similarly with the inclination to eat when one feels hungry. A well-designed rational agent needs however to be able to override such Q&I modules if reasoning indicates that it would be better to do so.

As Pollock himself points out (1995: 34-35), all kinds of evaluative attitudes other than situation-likings are subject to evaluative rational criticism, i.e., to criticism which is not a criticism of any beliefs on which they rest. Instrumental desires can be criticized by evaluating the plan from which they are derived. Primi-

tive 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.

Further, as Pollock also points out, not all reasoning is epistemic; here he explicitly dissents from Hume. 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.

6. Weaknesses of Pollock's BDIL model

An obvious objection to Pollock's model is that it requires a cardinal measure of situation-likings. While one can assign such numbers to a computational simulation of a rational agent, human beings clearly do not consciously associate with their awareness of their present situation some number which measures how much they like it. Pollock does suppose, quite plausibly I think, that human beings have a "quantitative feel" (1995:17) for how much they like a given situation which permits a certain comparative ordering. Consider any four token situations in which you have found yourself. Assign to them the labels "*a*", "*b*", "*c*" and "*d*" in such a way that you liked situation *a* more than you liked situation *b*, and you liked situation *c* more than situation *d*. Then you should be able to tell whether the first difference in liking was greater than, equal to or less than the second difference in liking; letting "*fx*" stand for "the likability of *x*", you might find that $(fa - fb) > (fc - fd)$. So far so good. But, in order to use such orderings as the basis for a cardinal measure of situation-likings, Pollock needs to make further rather complicated assumptions (Pollock 1995: 18 n. 13). It is doubtful whether such assumptions are justified.

Further, Pollock's model is incomplete in at least three important respects.

First, it is solipsistic, in the sense that there is no provision for verbal input from, or verbal output to, other autonomous rational agents, still less for back-and-forth discussion, whether argumentative or non-argumentative.

Second, it is egoistic, in that the function of the entire system is to make the world more to the liking of that system itself, without regard (except instrumentally) to whether its actions make the world more or less to the liking of other systems which have situation-likings and situation-dislikings. In calling Pollock's model egoistic, I do not mean to imply that the situation-likings which are at its basis have reference only to how the agent is faring. An agent might well find one situation

less likable than another only because someone else is worse off in the former situation. Most parents of small babies, for example, would find a situation more likable if the baby was healthy than if it was sick, quite apart from any inconveniences to themselves. But the relevance of the baby's situation to the practical reasoning of the parent is on Pollock's model a function only of the parent's likings. If the parent is indifferent between the health and the sickness of the baby, nothing in Pollock's account permits rational criticism of this indifference. It is in this sense that Pollock's model is egoistic. Morally speaking, Pollock's "rational agent" is a monster.

Third, Pollock's model is unsocial, in that his rational agent does not (and cannot) belong to any groups of autonomous rational agents with governance structures for making decisions about the actions of the group; it is a citizen of no country, belongs to no professional associations, owns no shares in any joint-stock company, has no immediate family, does not belong to a recreational bridge-playing group, etc.

A comprehensive theory of good practical reasoning would have to remedy all three of these lacks.

7. Conclusion

Pollock's model of practical reasoning has been computationally implemented in a comprehensive architecture for a rational agent which he calls OSCAR (Pollock 1999). His work illustrates a great advantage of computationally implementing philosophical theories: it brings to the fore new questions which were previously neglected. In the case of OSCAR, these include the need for Q&I modules, the necessity for a control structure for engaging in practical reasoning and the need to be able to override Q&I modules in the light of reflective reasoning. The need to design a system which combines epistemic and practical reasoning has produced a model of practical reasoning which is much more sophisticated and complex than anything previously produced. In particular, Pollock has made a strong case that practical reasoning requires not only the beliefs and desires which theorists of practical reasoning have required for millennia, and not just the additional distinct category of intentions for which Michael Bratman has argued, but also likings. And he has shown how a variety of transitions between mental states of these types are subject to rational criticism. At the same time, his model is incomplete in not allowing for communication between rational agents, social cooperation and the recognition of moral constraints. These three lacks are obviously interconnected.

Since the present paper was written, Pollock has developed a theory of rational decision-making for realistically resource-bounded agents (Pollock 2004). The theory is developed in three parts—on values, on probabilities, and on combining values and probabilities in decision-making. It has not been possible to incorporate in the present paper a description or evaluation of this theory. It appears to incor-

porate the model of practical reasoning described above, and thus to have the same strengths and weaknesses.

References

Aristotle (1984). *The Complete Works of Aristotle*, the revised Oxford translation edited by Jonathan Barnes. 2 vols. Princeton: Princeton University Press. First published ca. 355-322 BCE.

Bratman, Michael E. (1987). *Intentions, Plans and Practical Reason*. Cambridge, MA: Harvard University Press.

—. (1999). *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge: Cambridge University Press.

Hume, David (1975). *A Treatise of Human Nature*, edited by L. A. Selby-Bigge. Oxford: Clarendon Press. First published 1739, this edition first published 1888.

Mill, John Stuart (1888). *Utilitarianism*, 10th edition. London: Longmans Green. First edition published 1863.

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

—. (1999). *Download OSCAR*. http://www.u.arizona.edu/~pollock/oscar.html. Visited on April 24, 2003.

—. (2004). *Thinking about Acting: Logical Foundations for Rational Decision Making*. Pre-publication ms., 268 pages. Available until it is in press at http://oscarhome.soc-sci.arizona.edu/ftp/PAPERS/Thinking-about-Acting.pdf. Visited February 29, 2004.

David Hitchcock
Department of Philosophy
McMaster University
Hamilton, ON, Canada L8S 4K1

hitchckd@mcmaster.ca