Planning in OSCAR

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

The OSCAR project has two parallel goals —the formulation of a general theory of rationality, and the creation of a general-purpose automated reasoner implementing that theory. The theory of rationality takes as its starting point my own philosophical work in epistemology, probability, and philosophical logic. The current status of the OSCAR project is described in detail in my just-completed book *Cognitive Carpentry* (probably to be published by Bradford/MIT Press). At this point I have produced a general architecture for rational cognition, and an initial implementation of that architecture. The architecture constitutes an interest-driven defeasible reasoner that takes its data from various kinds of perceptual inputs, tries to answer questions posed by practical cognition, and uses those answers to direct its actions. The sense in which epistemic cognition is interest-driven is that practical cognition poses questions that epistemic cognition tries to answer, and the way in which epistemic cognition proceeds is a function, in part, of what questions it is trying to answer. The questions posed by practical cognition will be stored in a list of ultimate-epistemic-interests, along with instructions for what to do with answers. The underlying structure of the interest-driven reasoner is that propounded in Pollock [1990]. The basic idea is the very simple one that interest-driven reasoning proceeds forwards from the given information and backwards from the given interests. Defeasibility arises from the use of prima facie (defeasible) reasons in the arguments.. Prima facie reasons support their conclusions without logically entailing them. Associated with each prima facie reason is an array of "defeaters". Interests come in degrees, and reasoning is prioritized according to the degree of interest in the questions it is trying to answer. This general account is incorporated into the implemented epistemic reasoner.

I will take for granted a distinction between epistemic cognition (cognition about what to believe) and practical cognition (cognition about what to do), although I do not suppose that these can be entirely separated. The topic that interests me here is how to employ such a defeasible reasoner as the inference-engine in a rational agent capable of directing its behavior by planning. It is to be emphasized that my intention is not to produce new planning algorithms, but rather to produce a general architecture capable of treating the outputs of arbitrary planning algorithms in reasonable ways in a general context of defeasibly held beliefs.

At a very general level, it is easy to describe an architecture for plan-based practical reasoning. The reasoner must have mechanisms for (1) adopting goals, (2) initiating planning for the goals, (3) constructing candidate plans, (4) deciding what candidate plans to adopt, and (5) executing adopted plans.

For the purposes of this sketch, I will assume a simplistic view of goal adoption according to which the agent is equipped with a set of *optative dispositions* that, in response to beliefs about its environment, and possibly also nondoxastic environmental input, produce *desires*, which are states encoding goals. Desires have attached strengths, and priority is given to planning for stronger desires. So desires are put on a *desire-queue*, ordered by strength, and retrieved one at a time for planning initiation. Planning is an exercise in epistemic cognition. Whether a plan is likely to achieve a certain goal, and do so with certain expected costs, is a factual question of the sort with which epistemic cognition is intended to deal. Planning algorithms are thus viewed as part of epistemic cognition. In effect, they are viewed as intelligent schemes for variable instantiation in the attempt to find mathematical constructions (plans) satisfying specified constraints.

Once epistemic cognition has produced a plan, the agent must still decide whether to adopt it. This is not a trivial matter. For example, the agent might produce several different plans for achieving the same goal, and then it must decide which to adopt. Or the agent may produce several plans aimed at different goals, but also believe that the plans interfere with each other so that they should not all be adopted. Again, a choice must be made. The logic of these decisions turns out to be quite complicated. Plans are assigned *expected values*, but because plans can embed one another, we cannot simply require that adopted plans maximize expected value. Every plan can be embedded in another plan having a higher expected value. This is discussed in my [1992].

Once plans are adopted, *instrumental desires* are formed for the performance of nonbasic steps (steps that cannot be executed without further planning), and placed on the desire queue. These

give rise to hierarchical planning and subsidiary plans for how to perform those steps. Adopted plans are sent to the *plan executor*, which executes them by producing actions.

Plan-based practical reasoning is complicated by the need to make it defeasible. The treatment of plans must be defeasible for two separate reasons. First, the beliefs upon which planning is based are only held defeasibly. If we retract beliefs or adopt new beliefs, this may alter our assessment of whether a plan is likely to achieve its goals and of the cost of executing the plan, and in a rational agent, this may lead to an adopted plan being withdrawn. Second, the process of choosing plans for adoption must itself be defeasible. If two plans compete, and the second plan is better than the first, then the second plan should be adopted in preference to the first. In principle, there are always infinitely many possible plans for achieving a given set of goals. However, planning is computationally difficult, and an agent operating in real time does not have much time for planning, so it must make its decisions on the basis of very limited sets of candidate plans. Introspection suggests that humans frequently (maybe even generally) make their decisions on the basis of finding a single plan that is "good enough" without going on to search for potentially better plans. On the other hand, if they stumble upon a better plan later, they are able to acknowledge that and adopt it in the place of the earlier plan. In other words, the plan adoption itself was based upon the defeasible reason that the first plan was good enough, and subsequently defeated on the grounds that the second plan was better. (This is all a bit loose, but I have given a precise account in my [1992] and in *Cognitive Carpentry*.)

It is apparent that defeasibility is just as important in practical reasoning as in epistemic reasoning. This creates major problems. The structure of defeasible epistemic reasoning has proven very complicated, and the structure of the reasoner is even more so (see my [1992a]). Do we have to build a separate defeasible practical reasoner? There is every reason to expect that that would be just as complicated. Instead, I want to suggest a computational strategy for reducing defeasible practical reasoning to defeasible epistemic reasoning. Rather than requiring separate computational modules for defeasible epistemic reasoning and defeasible practical reasoning, human cognition makes do with a single module dedicated to epistemic reasoning, and then integrates practical reasoning into that module using a technical trick. The trick involves "doxastifying" normative judgments. Corresponding to the adoption of a plan is the "epistemic judgment" (i.e., belief) that *it should be an adopted plan.* This judgment is epistemic in name only. It requires no "objective fact" to anchor it or give it truth conditions. It is merely a computational device whose sole purpose is to allow us to use defeasible epistemic reasoning to accomplish defeasible practical reasoning. Let us abbreviate ⌜σ should be an adopted plan⌝ as ⌜σ is *adoptable*⌝. Then the defeasible reasoning underlying plan adoption can be translated into defeasible epistemic reasoning about plan adoptability, employing defeasible reasons like the following:

(E1) ⌜σ is a minimally good plan⌝ is a defeasible reason for ⌜σ is adoptable⌝.

(E2) ⌜α competes with σ and α is preferable to σ⌝ is a defeater for (E1).

The details of this defeasible epistemic reasoning will turn upon the account given of the logic of choosing between plans, and I will not go into to that here, but this gives the general idea.

Abbreviated Bibliography

Pollock, John L.

1986 *Contemporary Theories of Knowledge.* Rowman and Littlefield.

1987 Defeasible reasoning. *Cognitive Science* **11**, 481-518.

1989 *How to Build a Person; a Prolegomenon.* Bradford/MIT Press.

1990 Interest driven suppositional reasoning. *Journal of Automated Reasoning* **6**, 419-462.

1990a *Nomic Probability and the Foundations of Induction.* Oxford University Press.

1992 New foundations for practical reasoning. *Minds and Machines* **2**, 113-144.

1992a How to reason defeasibly. *Artificial Intelligence* **57**, 1-42.