Rationality in Decision Machines

Michael P. Wellman

University of Michigan, AI Laboratory
1101 Beal Avenue
Ann Arbor, MI 48109-2110
wellman@umich.edu

Abstract

This brief position paper[1] sketches a view, in many respects conventional, of the role of rationality in artificial intelligence. In this view, rationality is primarily a way to describe coherence between actions and agent attitudes such as belief and preference. *Decisions* are the fundamental unit of calculation for a rational computer, and therefore we properly refer to them as *decision machines*. Collections of interacting decision machines, or *decision factories*, should not in general be expected to be rational, although we should exploit the rationality of their constituent machines in configuring decision factories.

Machine Rationality

Computer scientists at least since Turing have understood that a program is an abstract specification of a *machine*, describable behaviorally in terms of the input/output relationship resulting from its computation. In adopting the common term, the field's pioneers probably meant to highlight the mechanical nature of computation, and perhaps also to emphasize the productive quality of running computer programs. The machine's *product* is of course the value of a function—the function represented by the program—at the point represented by the machine's input.

But often we find it helpful to view this product at a higher level, say, as the solution to some well-posed problem. Indeed, every subfield of computer science defines its own abstractions above the level of computing functions. For example, database specialists might prefer to view the input of their machine as a *query*, the product as *data*, and the machine's activity as *retrieval*. Of course we can always reduce this abstract database machine to the more concrete computing machine, but in doing so we may increase the description complexity and make it more difficult to analyze, predict, or design the behavior of database machines. In this case and most others, the reduction is perfect, in the sense that the lower-level machine exactly realizes the behavior of the abstract machine, and the relationship between them is typically described with complete precision.

Fifteen years ago, Newell proposed that a central characteristic of AI practice is a particular abstraction level at which we interpret the behavior of computing machines. Viewing a system at Newell's *knowledge level* entails attributing to the system knowledge, goals, and available actions, and predicting its behavior based on a *principle of rationality* that specifies how these elements dictate action selection. Rationality as applied here is a matter of coherence, defining a relation in which the knowledge, goals, and actions must stand. This is exactly the Bayesian view of rationality, in which knowledge and goals (or beliefs and preferences) are subjective notions, constrained only by self-coherence (consistency) and coherence with resulting behavior.

Newell acknowledged that the knowledge level is more radical than most other abstractions, in that we recognize that it is at best an approximation to the actual behavior described at a mechanistic (directly functional) level. That is, we employ the knowledge-level abstraction even though we lack a perfect reduction to systems described in purely computational terms. In fact, we know that such a reduction cannot exist, as the knowledge level (at least in its pure form) idealizes away the computational process itself.[2] Nevertheless, the knowledge-level perspective can be quite powerful, enabling us to predict the behavior of a system too complex or opaque to be described at a mechanistic level.

In introducing the knowledge-level idea, Newell proposed a particular basic rationality principle (along with several auxiliary principles):

> If an agent[3] has knowledge that one of its actions will lead to one of its goals, then the agent will select that action.

[1] World-wide web version (possibly more up-to-date) available at http://ai.eecs.umich.edu/people/wellman/decision-machine.html.

[2] For some good ideas about incorporating computational limits into machine rationality concepts, see the recent work of Russell and Subramanian (1995).

[3] The term Newell uses for the knowledge-level machine. In the remainder of this essay, the term *agent* is applied generically to denote the entity producing behavior.

This formulation can be criticized on several grounds, foremost of which is that it embodies, *logical* rather than *economic* rationality (Doyle, 1992). But let us put aside such concerns and consider the perspective that a rationality principle of this general form takes on computing machines. The relevant output identified by this principle is an *action selection*. Thus, action selections, or in effect, *actions*, are the products of a knowledge-level machine.

Actions and Decisions

Note that it is quite possible to view computing machines as producing actions, without accepting this particular rationality principle or any other part of the knowledge-level framework. In their comprehensive AI textbook, Russell and Norvig (1995) manage to cast most of the field as the attempt to build rational agents, where agents compute functions from perceptions to actions. *Knowledge-based agents* are a special case, where the agent's functional interface is in terms of declarative expressions. Actions are the answers to special query sentences, asking in effect "What is to be done?".

But exactly what are actions? Newell posits as a simplifying assumption that the agent's "body" (that is, its range of behavior) consists of a set of actions it can choose, acknowledging that in actuality the mode of interaction between an agent and its environment can be arbitrarily complex. Ultimately, actions are in the interpretation of a beholder. The true output of a computer program is a stream of physical encodings, which are then transduced by some external effectors into activity affecting the environment. But neither this activity nor the information states inducing it are inherently divisible in some particular way into discretely selectable action units (Rosenschein and Kaelbling, 1995).

There are at least two standard technical meanings of the term *action* in AI and the decision sciences. In one, typical of foundational treatments of decision theory, an act or action denotes the entire behavior of the agent in question. In the other, typical of AI planning discussions, an action is an atomic component (or step) of a more composite activity description, called a *plan*. An action by this latter interpretation is not satisfactory as the subject of a rationality criterion, however, because it is generally not possible to determine whether an individual action is rational separate from consideration of the entire plan of which it is part.[4] If it were, planning would be a trivial exercise.

An action by the former interpretation—or a plan by the latter—could be legitimately evaluated for rationality. This interpretation, however, does not yield a very practical rationality principle, because its unit of analysis is the entire behavior of the system. At a fundamental level, this is Savage's (1972) problem of *small worlds*. In Savage's words, the ideal rational agent

> …has only one decision to make in his whole life. He must, namely, decide how to live, and this he might in principle do once and for all.

Choice of this lifelong policy is the agent's *grand-world* decision problem. Decision scientists following (and including) Savage have rightfully deemed grand-world planning unrealistic, and instead focus on episodes which they assume can be reasonably isolated from the remainder of the agent's lifetime of activity. This is at best an approximation, and guaranteeing an approximation of acceptable quality may require considering episodes of impractically coarse granularity.

What we need is a more flexible unit of analysis—one that recognizes that the ultimate object for evaluation is the grand-world plan, but admits scrutiny of more localized computational products. Let Π be the *plan space*—the set of all grand-world plans available to the agent. We define a *decision* to be a commitment to a subset, Π', of Π. The choice of Π' over its complement constrains the activity of the agent, in any of a variety of ways.

For example, selection of an action (plan step) is a special case of a decision, in this case commitment to the class of plans containing that action, excluding those that do not. But decisions can also involve more complex patterns of activity, including conditional commitments, combinations of actions, or any expressible distinction in plan space, for that matter. Indeed, something of this nature is necessary to incorporate notions of intention (Pollack, 1992), since intentions are rendered superfluous if either individual actions or entire grand-world plans are adopted as the unit of analysis.

Since the most comprehensive rationality criteria are expressed in terms of grand-world plans, assessing the rationality of a decision is not entirely straightforward. To even entertain the possibility of rationality, we must attribute to an agent *choice*, that is, an ability to behave other than it actually behaves. In other words, we must view the computation as nondeterministic, or uncertain, at least from the observer's perspective. As a computation proceeds, uncertainty is progressively resolved, and its possible end results are narrowed. This is equivalent to saying that the agent *makes decisions*, committing to progressively smaller sets of possible plans.

In previous work I have proposed a general dominance-based criterion for assessing grand-world distinctions (Wellman, 1990), but more specialized approaches are appropriate for particular formulations of plan space. Of course, we can always revert to the special case of episodic decision making when that approximation is acceptable.

For present purposes, I wish to emphasize the flexibility of the concept of decision as defined here, and propose that decisions are the appropriate unit from which to formulate a comprehensive principle of rationality. By viewing our artifacts as *decision machines*, we focus on the relevant

[4] For some blatant illustrations of this, and an attempt to patch the problem, see Pollock (1992).

product—patterns of activity—at whatever granularity is appropriate.

Why Rationality?

It is emphasized repeatedly above that the essence of rationality is coherence, not competence. A coherence notion of some sort is *necessary* if we are to make any sense of our machine at the knowledge level. Otherwise, the attribution of knowledge and other agent attitudes to the machine would be completely unconstrained, therefore meaningless, therefore useless.

Hence the question "why adopt a rationality principle?" reduces to "why do we care to attribute knowledge and other so-called agent attitudes to our machines?". This latter question has been addressed by Newell, McCarthy, Dennett, Shoham, and others, and I will not attempt to reproduce their eloquence. Suffice it to say that the arguments all boil down to effectiveness of the rationality abstraction, where effectiveness is judged from the perspective of the observer or designer. Empirically, it appears that we can often effectively predict and explain the behavior of exceedingly complex or mechanistically ill-understood systems (such as people) by attributing to them agent attitudes such as beliefs, preferences, etc., … *and* rationality. The premise is that the complexity of sufficiently capable computing machines will also lie beyond the scope of program-level abstractions, and that the most effective way to deal with our decision machines will be in terms of beliefs, desires, intentions, and the like.

Even though the primary motivation for rationality is coherence, it is without a doubt advisable to consider competence when deciding *what* rationality principle to adopt. All else being equal, we would naturally prefer a rationality principle that is normatively compelling and comprehensive (such as Bayesian decision theory, in my view). The descriptive validity of alternative rationality principles for various classes of decision machines is an interesting and largely unexplored question. It would seem that absent countervailing evidence, we should aim to design machines adhering to the highest standards of rationality we can imagine. Relaxation of the standards should be entertained only for specifically motivated compromises.

Decision Factories

Let us conclude by turning briefly to the question, "What should we expect from a *collection* of interacting decision machines?" From rationality of the individuals does not follow rationality of the collection, and an attempt to achieve rationality at the aggregate level may well be misguided. Designers of *decision factories* should account for and exploit—via *incentive engineering*—rationality of the component machines in designing the system configuration, but should not expect in general that coherent single-agent attitudes can be attributed to the aggregate system. If it is somehow crucial that rationality be achieved overall, then perhaps it should be sacrificed in the components.

If one is seeking ideas about collections of rational agents, my summary suggestion is to look toward microeconomics (Wellman, 1995). Indeed, its pervasive use of the rationality abstraction is one of the main features distinguishing economics from other social science disciplines. In that sense economics and artificial intelligence are natural allies, and we could gain much by adapting the concepts, mechanisms, and theories of economics for the design of decision factories. For some approaches that do this, see Clearwater (1995), Rosenschein and Zlotkin (1994), and Wellman (1993).

Acknowledgments

This work has been supported in part by an NSF National Young Investigator award.

References

Clearwater, S. H., ed. (1995). *Market-Based Control: A Paradigm for Distributed Resource Allocation*. World Scientific.

Doyle, J. (1992). Rationality and its roles in reasoning. *Computational Intelligence* **8**: 376-409.

Newell, A. (1982). The knowledge level. *Artificial Intelligence* **18**: 87–127.

Pollack, M. E. (1992). The uses of plans. *Artificial Intelligence* **57**: 43-68.

Pollock, J. L. (1992). New foundations for practical reasoning. *Minds and Machines* **2**: 113–144.

Rosenschein, J. S., and G. Zlotkin (1994). *Rules of Encounter: Designing Conventions for Automated Negotiation among Computers*. MIT Press.

Rosenschein, S. J., and L. P. Kaelbling (1995). A situated view of representation and control. *Artificial Intelligence* **73**: 149-173.

Russell, S. J., and P. Norvig (1995). *Artificial Intelligence: A Modern Approach*. Prentice-Hall.

Russell, S. J., and D. Subramanian (1995). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research* **2**: 575-609.

Savage, L. J. (1972). *The Foundations of Statistics*. Second edition. New York, Dover Publications.

Wellman, M. P. (1990). *Formulation of Tradeoffs in Planning Under Uncertainty*. London, Pitman.

Wellman, M. P. (1993). A market-oriented programming environment and its application to distributed multicommodity flow problems. *Journal of Artificial Intelligence Research* **1**: 1-23.

Wellman, M. P. (1995). The economic approach to Artificial Intelligence. *ACM Computing Surveys* (to appear).