Rationality and its Roles in Reasoning
(Extended Abstract)

Jon Doyle*
Massachusetts Institute of Technology
Laboratory for Computer Science
545 Technology Square
Cambridge, Massachusetts 02139, USA

Abstract

The economic theory of rationality promises to equal mathematical logic in its importance for the mechanization of reasoning. We survey the growing literature on how the basic notions of probability, utility, and rational choice, coupled with practical limitations on information and resources, influence the design and analysis of reasoning and representation systems.

Introduction

People make judgments of rationality all the time, usually in criticizing someone else's thoughts or deeds as irrational, or in defending their own as rational. Artificial intelligence researchers construct systems and theories to perform or describe rational thought and action, criticizing and defending these systems and theories in terms similar to but more formal than those of the man or woman on the street.

Two conceptions of rationality dominate these judgments: a *logical* conception used to judge thoughts, and an *economic* one used to judge actions or choices. For example, when people criticize as irrational someone who asserts both a proposition p and its contrary $\neg p$, or who asserts p and $p \Rightarrow q$ but refuses to accept q, they refer to a logical sense of rationality. Correspondingly, when some people criticize others for irrationally wasting their money on state lotteries, in which the predictable result of prolonged gambling aimed at winning money is, in fact, to lose money, the critics have in mind the economic sense of rationality.[1] In classical terms, logic concerns Truth, while economics concerns Goodness (though a case can be made that neither says much about either).

Traditionally, much work in artificial intelligence has been greatly swayed by the "logicist" view that logic is the theory of the ideal good thinking desired of all intelligent agents—in particular, that beliefs should be consistent and inferences sound—and has paid much less attention to the economic sense of rationality. One may interpret much non-logicist work on heuristics as implicitly concerned with rationality in the economic sense, but little of this work discusses rationality explicitly or employs any of the formal tools offered by the mathematical theory of rational choice. Recently, however, interest in economic rationality and its formal theory has grown as researchers have sought to find methods for reasoning under uncertainty, for controlling reasoning, and for putting heuristic methods on sound theoretical bases—each one an issue on which logic alone provides little guidance.

The purpose of this paper is to introduce the basic notions of economic rationality. These constitute a rich set of conceptual and mathematical tools for analyzing information and behaviors, and provide the proper framework for addressing the problem of how one should think, given that thinking requires effort and that success is uncertain and may require the cooperation of others. Though it provides an attractive ideal, however, the level of information and computational ability demanded by the theory render straightforward applications of the theory impractical, as was pointed out early on by Simon [1955], who introduced the term *bounded* rationality (also called *limited* rationality) for the rationality that limited agents may feasibly exhibit.

We first summarize the basic concepts of economic rationality and identify the principal roles economic rationality plays in the theory and practice of artificial intelligence. The highly abbreviated remainder of this extended abstract examines various impediments to achieving rationality, indicates recent developments on techniques for reasoning rationally in the presence of these limitations, and points out some future directions for research.

Economic rationality

The fundamental issue in the theory of economic rationality is *choice among alternatives*. Economic rationality simply means making "good" choices, where

*This work was supported by National Institutes of Health Grant No. R01 LM04493 from the National Library of Medicine.

[1] Gambling may be rational if the gambler also has non-monetary aims, such as entertainment.

goodness is determined by how well choices accord with the agent's *preferences* among the alternatives. We summarize the elements of this theory here: for more complete expositions, see [Debreu, 1959; Savage, 1972; Jeffrey, 1983].

Preference

The notion of preference is the fundamental concept of economic rationality. We write $A \prec B$ to mean that the agent prefers B to A, and $A \sim B$ to mean that the agent is *indifferent* between the two alternatives, that is, considers them equally desirable or undesirable. We also write $A \precsim B$ (B is *weakly preferred* to A) to mean that either $A \sim B$ or $A \prec B$. The collection of all these comparisons constitutes the agent's set of preferences.

Rational agents choose *maximally preferred* alternatives. If $\mathcal{A} = \{A_1, \ldots, A_n\}$ is the set of alternatives, then A_i is a rational choice from among these alternatives just in case $A_j \precsim A_i$ for every $A_j \in \mathcal{A}$. It is not required that the agent explicitly calculate or compute the maximality of its choices, only that it chooses alternatives that are in fact maximal according to its preferences. There may be several rational choices, or none at all if the set of preferences is inconsistent or the set of alternatives is infinite.

The theory requires, as a minimum basis for rationality, that strict preference is a strict partial order, indifference is an equivalence relation, indifference is consistent with strict preference, and any two alternatives are either indifferent or one is preferred to the other. More succinctly, weak preference is a complete preorder, or formally, for all alternatives A, B, and C:

1. Either $A \precsim B$ or $B \precsim A$, *(completeness)*
2. If $A \precsim B$, then $B \not\prec A$, and *(consistency)*
3. If $A \precsim B$ and $B \precsim C$, then $A \precsim C$. *(transitivity)*

These rationality constraints ensure that there is always at least one rational choice from any finite set of alternatives.

Utility

The rationality constraints imply that we may represent the set of preferences by means of a numerical *utility* function u which ranks the alternatives according to degrees of desirability, so that $u(A) < u(B)$ whenever $A \prec B$ and $u(A) = u(B)$ whenever $A \sim B$. By working with utility functions instead of sets of preferences, we may speak of rational choice as choosing so as to maximize utility. Any strictly increasing transformation of a utility function will represent the same set of preferences, and will provide the same choices under maximization.

The distinction between the (monetary or computational) costs or values of something and its utility or disutility is one of the great strengths of the theory of economic rationality, as compared with everyday accounting. Utility is an aggregate measure of all dimensions of worth, relative to the agent and the agent's situation, and mere costliness is no guarantee of utility. In reasoning, for example, the utility of some conclusion usually depends on numerous variables: on what the question was, in what circumstances it was asked, on how it is to be used, on when the answer was obtained, and on how reliable the conclusion is. Whether deriving the conclusion is easy or hard does not affect these factors.

One cannot define the notions of preference and utility purely in terms of beliefs and goals, for these are independent notions. Goals only state what is desired, and do not give any information about the relative merits of different desirable alternatives (see [Dean and Wellman, 1989]).

Decision theory

Most work in artificial intelligence that makes use of economic rationality draws on the specific theory of *subjective Bayesian decision theory* [Savage, 1972], hereafter simply called decision theory. Compared with the basic theory, decision theory adds probability measures p_A which indicate the likelihood of each possible outcome for each alternative $A \in \mathcal{A}$. Decision theory also strengthens the notion of utility from an *ordinal* utility function u to a *cardinal* utility function U. Ordinal utility functions use numerical values simply as ways of ranking the alternatives in a linear order. It does not make sense to say that an ordinal utility of 10 is twice as good as an ordinal utility of 5, any more than it makes sense to say that the tenth smallest person in a room is necessarily twice as tall as the fifth smallest. Amounts of cardinal utility, in contrast, can be added and subtracted to produce other amounts of utility. This makes it possible to combine the utilities foreseen in different possible outcomes of A into the *expected utility* $\hat{U}(A)$, defined to be the utility of all possible outcomes weighted by their probability of occurrence, or formally,

$$\hat{U}(A) \stackrel{\text{def}}{=} \sum_S p_A(S)U(S), \qquad (1)$$

where the sum (more generally, an integral) ranges over all possible situations or states of nature under discussion. For example, if there are exactly two possible states S_1 and S_2 of respective utilities 5 and 7, and if the respective probabilities of these states obtaining as outcomes of alternative A are .1 and .9, then the expected utility of A is just $.1(5) + .9(7) = 6.8$. The decision-theoretic definition of preference is then

$$A \precsim B \text{ if and only if } \hat{U}(A) \leq \hat{U}(B).$$

Like the theory of preference, the axioms for decision theory involve qualitative orderings $\precsim_p$ and $\precsim_U$ of outcomes according to relative likelihood and desirability.

These comparisons we also call beliefs and preferences (distinct from overall preferences or judgments).

Just as cost is not the same as utility, expected utility is not the same as average cost, even when utility is a function of cost alone. Expected utility necessarily averages over utilities, not over the variables on which utilities depend. For example, bicycles designed to fit the average size rider perfectly serve an evenly mixed population of tall adults and short children only poorly. In the same way, expected computational utility need not be a function of average running time.

The need for economic rationality

Logical and economic notions of rationality can be used either *descriptively*, as giving sets of concepts and mathematical tools with which reasoning and action may be formalized and analyzed, or *normatively*, as giving standards of correctness to which reasoning and action must conform. Descriptively, for example, logic has been used to formalize beliefs and other representations, to determine what hypotheses are possible given the reasoner's beliefs, and to determine which methods have any possibility of achieving specified goals. Similarly, economic rationality may be used descriptively to identify the conditions under which one inference technique is better than another, or to explain why a technique is good or bad in specific circumstances. In particular, the theory may be applied in AI to provide a formal analysis of informally developed techniques (e.g., [Langlotz *et al.*, 1986]). Normatively construed, however, logical and economic rationality are at odds with one another over how one should think. We begin by examining the normative use of logic, which we will call *logicism*.

Logicism

The logicist school of artificial intelligence views reasoning as a form of logical inference and seeks to construct general-purpose deduction systems in which axioms state what is held true and goals state what is desired to be proven true (or to be achieved as the result of actions). Logicism's standard asks whether the reasoner's beliefs are consistent and inferences sound (and sometimes whether the beliefs and inferences are complete as well).

Logicism is not a complete theory of thinking by itself, since it views the problem of how reasoning should be conducted as a pragmatic question outside the realm of the theory of thinking proper. In logic, any consistent set of beliefs and any sound inference is as good as any other, and the only guidance logicism seems to offer the reasoner is the rule

If it's sound, do it!

Logicism ignores issues of the *purpose* of reasoning (other than to suppose the existence of externally posed goals) and of the value of beliefs and inferences to the reasoner, basing inferences purely on the logical form of the beliefs and goals. It ignores questions of whether the reasoner should or should not draw some inference, and whether one inference is better or more appropriate than another. The usual result is that purely logical reasoners make many worthless inferences, since sound worthwhile inferences may be of the same logical form as sound worthless inferences (cf. [McDermott, 1987]).

Making worthless inferences would not matter if reasoners were not expected to arrive at conclusions and take actions within appropriate lengths of time. But most reasoning does have some temporal purpose. To reason intelligently, the reasoner must know something about the value of information and about which methods for achieving goals are more likely to work than others, and must prudently *manage* the use of its knowledge and skills by taking into account its own powers, limitations, and reliability (cf. [Doyle and Patil, 1989]). For example, for some questions it may be clear that no answer is possible, or that finding the answer will take too long, in which case the reasoner may conclude "I don't know" right away. This might save enough time for the reasoner to successfully answer other questions. Alternatively, the exact answer might appear to take too long to determine and the reasoner may choose to look for an adequate approximate answer that can be found quickly. In either case, the reasoner performs better by anticipating limits and reasoning accordingly than by simply suffering limits. Simply deducing new conclusions until reasoning is terminated by reaching an answer or a deadline leads to haphazard performance, in which the reasoner succeeds on one problem but fails on seemingly identical ones that few people would distinguish, with no discernible pattern to help predict success or failure.

Heuristic problem solving

Many non-logicist approaches to AI also downplay issues of making rational choices. For example, in his characterization of the knowledge level, Newell [1982] formulates what he views as the fundamental principle of rationality as follows:

> "*Principle of rationality.* If an agent has knowledge that one of its actions will lead to one of its goals, then the agent will select that action." [Newell, 1982, p. 102]

(Cf. Cherniak's [1986] principles of "minimal rationality.") Newell calls this principle the "behavioral law that governs an agent, and permits prediction of its behavior". Since this principle ignores comparisons among goals and among the different methods for goals, the activity it prescribes is almost as indifferent as in the logicist rule above. Many AI systems ignore issues of choice in much the way the principle suggests. Newell eventually adds auxiliary principles about how to act given multiple goals and multiple methods, and acknowledges that these ultimately lead to the economic

theory of rationality, but nevertheless bases his theory of knowledge on this fundamental principle alone.

Newell's principle of rationality notwithstanding, many in artificial intelligence, including Newell, have long recognized the limitations of unguided reasoning and have advanced the notion of *heuristics* as central to effective problem solving. Heuristics usually amount to holding beliefs or making inferences that are deemed to be useful though sometimes unsound or mistaken. Indeed, the standard guideline in heuristic problem solving is the rule

If it seems useful, do it!

But the notion of usefulness motivating the use of heuristics has rarely been made formal, which has brought much work on heuristic methods into disrepute among logicians, mathematicians, and formally-minded AI theorists. To many, lack of any respectable alternative has been the main attraction of logicism as a formal theory of thinking.

Economic rationality

Economic rationality provides an answer both to the problem of controlling reasoning and to the informality of heuristics. In the first place, *economic rationality is the proper standard for the knowledge level* (cf. Baron's [1985] psychological perspective). It adds formal theories of utility and probability to the logicist formulation of belief and inference, and provides a new norm for guiding reasoning and action, namely that the reasoning activities performed should have maximal expected utility among those open to the agent. It subsumes portions of the logicist approach, since logic can be viewed as the theory of *certain* beliefs, that is, beliefs of probability 1, and the axioms of probability require that certain beliefs be consistent just as in logicism.[2] It also decouples the notion of rationality from the notion of intelligence. Intelligence depends on the actual knowledge possessed and used, while rationality merely depends on possession and use of types of knowledge, namely expectations and preferences. Secondly, heuristics may be formalized as methods for increasing the expected utility of reasoning. Since different alternatives may be open to the agent at different times in different reasoning processes, the task for artificial intelligence is to examine each of these situations and determine both the possible methods and their relative utilities.

Instead of competing as normative theories, logical and economic notions of rationality fill complementary needs. Logic serves to describe the possibilities for reasoning and action, while economics serves to prescribe choices among these. Logic plays a descriptive role in developing formulations of problems; economics plays a normative role in choosing both the problems to solve and the means of solving them.

Rationality in limited agents

The normative use of decision theory provides a standard for rationality, but one which is often unattainable due to limitations on the available information or resources. For example, beliefs and preferences may be incomplete in that a reasoner may not know whether one circumstance is more likely than another or which it prefers, or it may not know all the consequences of its beliefs and preferences (that is, it may not be logically omniscient). Beliefs and preferences may be inconsistent due to conflicts among authorities or commonsense theories, or indeterminate if they vary from situation to situation or are derived from information distributed among different frames, perspectives, or subagents (Thomason [1986] calls this "context sensitivity"). If beliefs and preferences cannot be revised quickly enough to account for new information, they also may exhibit inertia.

Similarly, computational resources, such as the time available for making decisions or the space available for representing information, may be limited. Some non-computational resources, such as the effort or cooperation required of human knowledge engineers, expert informants, or end users may be even more limited than computational resources. In addition, some limitations stem from the reasoner's architecture or organization itself, since the importance of particular time and space limitations depends on the structures and operations provided by the architecture, which determine the costs and reliabilities of reasoning.

While informational, resource, and organizational limitations may all be subject to change by progress in science and technology, there may also be physical and metaphysical limitations not subject to human influence. The known physical limitations, for example, include the speed of light, the interference of simultaneous measurements, finiteness of the matter and energy available to represent information, and the inertia of matter and energy.

The metaphysical limitations concern whether rationality is well defined or even possible in principle. Milnor [1954], for example, listed a number of intuitively desirable properties of rational decisions under uncertainty, each of which is satisfied by some extant theories, and then proved the set of these properties to be inconsistent. This suggests that there may be several essentially different intuitions underlying the notion of rationality (cf. [Touretzky *et al.*, 1987]). There is also strong psychological evidence that expected utility does not capture a realistic notion of preference.

[2] Of course, the theory of economic rationality itself may be axiomatized as a logical theory, just like any other theory (e.g., meteorology). This does not mean that the notions of logic subsume those of economical rationality (resp. meteorology), since then logic supplies only the form of the theory. In contrast, logicism uses logic to supply the content of the theory of thinking.

Preference is not linearly additive in the probabilities of events as is required in equation (1), and humans often exhibit preference reversals and so-called framing and anchoring effects (see [Machina, 1987; Kahneman *et al.*, 1982]). Finally, it may be that overall utility functions simply do not exist. The existence of a utility function entails that all values can be combined into a single scale of desirability, and this may not always be possible [Van Frassen, 1973; Nagel, 1979; Doyle and Wellman, 1989].

All these limitations mean that the rationality exhibited by limited agents will be somewhat different from the rationality presumed in the idealizations of decision theory. Rationality in the ideal theory considers only whether the *results* of choices best satisfy the agent's preferences, while rationality in limited agents also considers whether the agent makes good choices in the *process* of deciding how to apply its efforts in reasoning toward a decision. Rationality when the costs of deliberation are taken into account is called "Type 2" rationality by Good [1971] and "procedural" rationality by Simon [1976], as opposed to "Type 1" or "substantive" rationality in which the costs of reasoning are ignored. What is rational for one agent may be in direct conflict with what is rational for agents with different (or no) limitations. This is clearest in the play of chess, where increasing search can successively reveal new threats and new benefits, possibly leading the reasoner to vacillate about whether some move is good or bad as the time available for searching increases.

Achieving Type 2 or procedural rationality means optimizing the overall degree of rationality by making rational choices about what inferences to perform, which methods to apply, and how (or how long) to apply them. Agents that recognize their own limitations and purposes and guide their actions and reasoning accordingly exhibit much of what the Greeks called *sophrosyne*, that is, temperance or self-control. But it does not always make sense to think a lot about how to think. That is, if the point of guiding reasoning is to arrive at desired conclusions more quickly, extensive calculations about which inference to draw at each step may consume more time than they save. Rational guidance of reasoning thus requires striking a balance between control computations and reasoning computations. The proper balance is, of course, found by choosing the amount of time to spend on control computations rationally so as to achieve the best performance.

Making control decisions rationally raises the problem of infinite regress, since trying to control the cost of making rational control decisions by means of additional rational control decisions creates a tower of deliberations, each one concerned with the level below (see [Doyle, 1980; Lipman, 1989]). In practice, the deliberative information available at higher levels but unavailable at lower ones decreases rapidly as one ascends the reflective tower, and most systems rely on well-chosen default choices at the first or second levels instead of long episodes of reflection upon reflection. In theory, halting deliberation at one level amounts to making the decisions for all higher levels at once, and rationality in this setting would seem to mean that the halting point can be judged rational after the fact, that is, as rational given all the expectations and preferences that result from making all these decisions at once. Rawls [1971] calls this condition *reflective equilibrium*. Jeffrey [1983] calls such decisions *ratified* decisions.

Specific roles for rational choice

AI has developed many apparently useful techniques for reasoning and representation, such as depth-first and A* search, dependency-directed backtracking, constraint propagation, explanation-based learning, etc. Considerable insight might be gained by analyzing these theoretically and empirically in economic terms, both to compare alternative methods with each other, and to find the conditions under which individual techniques and representations increase (or decrease) expected utility. Most heuristic methods are thought to increase utility, but at present most are used without any real information about their probability of usefulness. Indeed, users are sometimes warned that one must have substantial experience with some techniques just to be able to tell when using the techniques will help rather than hurt. Can we demonstrate that these expectations of heuristic value are reasonable? More generally, can we make precise the assumptions about probabilities and utilities that underlie these judgments?

We here enumerate some tasks in which rational choice would seem to play a significant role. (See the full version of this paper for a more complete discussion.) Substantial work has already been done on some of these, but others have seen only initial explorations.

Rational approximations: Reasoners uncertain about the time available may employ approximation methods in which the expected utility or the probability of correctness of partial answers increase monotonically as further effort is applied. These include *flexible computations* or *anytime algorithms* [Horvitz, 1988; Dean and Boddy, 1988], and *probably approximately correct* (PAC) algorithms [Valiant, 1984].

Rational assumptions and belief revision: Default rules may be viewed as implicitly rational decisions about reasoning [Doyle, 1983; Doyle, 1989; Langlotz and Shortliffe, 1989; Shoham, 1988]. More generally, the reason for recording a belief in memory is the expectation that it will be useful in future reasoning and that the effort needed to rederive or replace it outweighs the utility of the memory resources consumed by recording it. It should be removed from memory only if it is expected to undermine the efficacy of actions enough to justify the effort of removing it.

Theories of *conservative* belief revision typically adopt principles like minimizing the number or set of

changed beliefs [Harman, 1986; Gärdenfors, 1988]. But these principles do not take into account any of the reasoner's preferences among different possible revisions, which means that revisions may be less rational than necessary. Rational choice would be especially valuable in the special case of backtracking, both in choosing which assumptions to abandon, and more fundamentally, in deciding whether analyzing and removing the inconsistency will be worth the effort.

Rational representations of inconsistent information: Theories of inheritance and nonmonotonic logics give the appearance of consistency by means of *credulous* representations, in which maximal consistent subsets are used to *represent* (in the sense of [Doyle, 1988; Doyle, 1989]) inconsistent sets of rules, and *skeptical* representations, in which the intersection of all maximal consistent subsets represents the inconsistent information [Touretzky *et al.*, 1987]. Neither skepticism nor credulity is rational in all situations, and choosing an appropriate representation may be difficult [Doyle and Wellman, 1989].

Rational search and inference: Russell and Wefald [1989] have developed explicit formulas and estimation techniques for decision-theoretic control of A* and other search methods (see also [Etzioni, 1989; Hansson and Mayer, 1989]). Horvitz [1988] and Breese and Fehling [1988] have examined control of larger-scale computational methods, while Smith [1986] has developed techniques for estimating the costs of deductive inference methods. See [Dean, 1990] for a detailed survey of this area.

Rational learning: Indiscriminate memorization leads to clogging memory with records of dubious value [Minton, 1990]. Thus it is important to make rational decisions about what to memorize, what to forget, what information to summarize, and what summaries to memorize, as well as how carefully or precisely to classify or categorize new objects in taxonomies and how to most efficiently organize taxonomies is very important.

Rational planning: Wellman [1990] describes a planner which uses *dominance* relations among plans to guide the search and to isolate the fundamental tradeoffs among methods, tradeoffs that remain valid even if the details of the situation change. Expected utility also provides a criterion for deciding whether to plan for contingencies.

Other tasks

Finding good representations for probabilities and preferences (see [Horvitz *et al.*, 1988; Pearl, 1988; Wellman, 1990]) would enable more rapid progress on the specific applications of rational choice discussed above. Other tasks include the following (more can be found in the full version).

Automate decision formulation and analysis: The field of *decision analysis* [Howard and Matheson, 1984; Raiffa, 1968] bears striking similarities to the more recent efforts in AI on developing expert systems [Horvitz *et al.*, 1988]. Initial efforts have been made towards automating the formulation of decisions (see [Breese, 1987; Wellman, 1990]), and at providing automatic tools to assist human analysts [Holtzman, 1989; Wellman *et al.*, 1989].

Exploit economic theory: Economics has substantial theories of common ways of organizing human societies or businesses that have received only initial exploration in AI, such as markets [Huberman, 1988] and more general social decision-making frameworks mixing authority relationships with decisions among equals [Minsky, 1986]. AI needs to exploit these theories, and especially the growing work in economics on applying economic models to modeling the attitudes and mental organization of the individual agent (see, for example, [Schelling, 1980; Thaler and Shefrin, 1981]).

Design provably rational architectures: Much might be learned by attempting to design general architectures for rational reasoning and action that are structured so as to permit clean theoretical analyses of their fundamental properties.

Reform AI education: The practice of teaching AI without prerequisites beyond elementary computer science is becoming increasingly untenable. There are now substantial theoretical foundations for portions of artificial intelligence, including both the basics of modern logic and the basics of economics and decision theory. Students intending serious study of AI need exposure to these foundations through courses in elementary logic and basic decision analysis, and possibly the foundations of decision theory and microeconomics as well. Simply including a couple lectures in an introductory AI class is probably not adequate.

Conclusion

Artificial intelligence has traveled far under the power of two ideas: exploiting logical inference as a method of reasoning, and using informal heuristics to direct reasoning toward useful conclusions. We have some understanding of systems based on logical inference, but making further progress toward flexible and intelligent reasoners requires understanding the capabilities and behaviors of systems guided by heuristics. Obtaining such understanding will be difficult without ways of analyzing, characterizing, and judging heuristics in terms as precise as those of logic. Fortunately, the economic theory of rational choice offers formal tools for understanding heuristics and other methods of guiding reasoning. In fact, economic rationality appears to offer a much-needed knowledge-level standard for how one should think, rather than simply enumerating ways in which one might think.

In spite of its attractions as a precise standard for reasoning and action, the theory of rational choice cannot be adopted uncritically for two reasons. First of all,

it places unreasonable demands on the knowledge and inferential abilities of the reasoner. Second, it is, like logic, a purely formal theory, and says nothing specific about what reasoning is actually useful. Applying rationality to reasoning and representation thus requires formulating realistic measures of cognitive utility, obtaining realistic expectations about the effects of reasoning, and developing cost-effective mechanisms for combining this information. Many fundamental and practical difficulties remain, but there is no alternative to facing them. If AI is to succeed, the issues of expectations, preferences, and utility cannot be ignored, and even using a problematic theory of rationality seems more edifying than using logic and informal heuristics alone.

In summary, logic and economic rationality are not competing theories, but instead are two complementary parts of the solution. Logic provides ways of analyzing meaning and possibility, while economics provides ways of analyzing utility and probability. We need to investigate how to integrate these theories in useful ways that recognize that meaning, possibility, utility, and probability must all be evaluated with respect to changing purposes and circumstances.

Acknowledgments

I wish to thank Ramesh Patil, Peter Szolovits, and Michael Wellman for reading drafts, Rich Thomason for lending me some of his notes, and Tom Dean, Othar Hansson, Eric Horvitz, Barton Lipman, Andrew Mayer, Stuart Russell, Joseph Schatz, and David Smith for valuable discussions.

References

[Baron, 1985] J. Baron. *Rationality and Intelligence.* Cambridge University Press, Cambridge, 1985.

[Breese, 1987] J. S. Breese. Knowledge representation and inference in intelligent decision systems. Research Report 2, Rockwell International Science Center, Palo Alto, 1987.

[Breese and Fehling, 1988] J. S. Breese and M. R. Fehling. Decision-theoretic control of problem solving: Principles and architecture. *Proc. Fourth Workshop on Uncertainty in AI*, pp. 30–37, 1988.

[Cherniak, 1986] C. Cherniak. *Minimal Rationality.* MIT Press, Cambridge, 1986.

[Dean, 1990] T. Dean. Decision-theoretic control of inference for time-critical applications. Brown University, Providence, RI, 1990.

[Dean and Boddy, 1988] T. Dean and M. Boddy. An analysis of time-dependent planning. *Proc. AAAI-88*, pp. 49–54, 1988.

[Dean and Wellman, 1989] T. Dean and M. P. Wellman. On the value of goals. *Proc. Rochester Planning Workshop: From Formal Systems to Practical Systems*, pp. 129–140, 1989.

[Debreu, 1959] G. Debreu. *Theory of Value: an axiomatic analysis of economic equilibrium.* Wiley, New York, 1959.

[Doyle, 1980] J. Doyle. A model for deliberation, action, and introspection. Report AI-TR-581, MIT AI Lab, 1980.

[Doyle, 1983] J. Doyle. Some theories of reasoned assumptions: an essay in rational psychology. TR 83-125, Department of Computer Science, Carnegie Mellon University, 1983.

[Doyle, 1988] J. Doyle. Artificial intelligence and rational self-government. TR CS-88-124, Carnegie-Mellon University Computer Science Department, 1988.

[Doyle, 1989] J. Doyle. Constructive belief and rational representation. *Computational Intelligence*, 5(1):1–11, February 1989.

[Doyle and Patil, 1989] J. Doyle and R. S. Patil. Two dogmas of knowledge representation: language restrictions, taxonomic classifications, and the utility of representation services. Report TM-387b, MIT Laboratory for Computer Science, 1989.

[Doyle and Wellman, 1989] J. Doyle and M. P. Wellman. Impediments to universal preference-based default theories. *Proc. First Int. Conf. on Principles of Knowledge Representation and Reasoning*, pp. 94–102, 1989.

[Etzioni, 1989] O. Etzioni. Tractable decision-analytic control. *Proc. First Int. Conf. on Principles of Knowledge Representation and Reasoning*, pp. 114–125, 1989.

[Gärdenfors, 1988] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* MIT Press, Cambridge, MA, 1988.

[Good, 1971] I. J. Good. The probabilistic explication of information, evidence, surprise, causality, explanation, and utility. In V. P. Godambe and D. A. Sprott, eds., *Foundations of Statistical Inference*, pp. 108–127. Holt, Rinehart and Winston, Toronto, 1971.

[Hansson and Mayer, 1989] O. Hansson and A. Mayer. Heuristic search as evidential reasoning. In *Proc. Fifth Workshop on Uncertainty in AI*, pp. 152–161, 1989.

[Harman, 1986] G. Harman. *Change in View: Principles of Reasoning.* MIT Press, Cambridge, MA, 1986.

[Holtzman, 1989] S. Holtzman. *Intelligent Decision Systems.* Addison-Wesley, Reading, MA, 1989.

[Horvitz, 1988] E. J. Horvitz. Reasoning under varying and uncertain resource constraints. *Proc. AAAI-88*, pp. 111–116, 1988.

[Horvitz *et al.*, 1988] E. J. Horvitz, J. S. Breese, and M. Henrion. Decision theory in expert systems and

artificial intelligence. *J. Approximate Reasoning*, 1988.

[Howard and Matheson, 1984] R. A. Howard and J. E. Matheson, eds. *The Principles and Applications of Decision Analysis.* Strategic Decisions Group, Menlo Park, CA, 1984.

[Huberman, 1988] B. A. Huberman. *The Ecology of Computation.* North-Holland, Amsterdam, 1988.

[Jeffrey, 1983] R. C. Jeffrey. *The Logic of Decision.* University of Chicago Press, Chicago, 2nd ed., 1983.

[Kahneman *et al.*, 1982] D. Kahneman, P. Slovic, and A. Tversky, eds. *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge University Press, Cambridge, 1982.

[Langlotz and Shortliffe, 1989] C. P. Langlotz and E. H. Shortliffe. Logical and decision-theoretic methods for planning under uncertainty. *AI Magazine*, 10(1):39–47, 1989.

[Langlotz *et al.*, 1986] C. P. Langlotz, E. H. Shortliffe, and L. M. Fagan. Using decision theory to justify heuristics. *Proc. AAAI-86*, pp. 215–219, 1986.

[Lipman, 1989] B. Lipman. How to decide how to decide how to ...: Limited rationality in decisions and games. *AAAI Symp. on AI and Limited Rationality*, pages 77–80, 1989.

[Machina, 1987] M. J. Machina. Choice under uncertainty: Problems solved and unsolved. *J. Economic Perspectives*, 1(1):121–154, Summer 1987.

[McDermott, 1987] D. McDermott. A critique of pure reason. *Computational Intelligence*, 3:151–160, 1987.

[Milnor, 1954] J. Milnor. Games against nature. In R. M. Thrall, C. H. Coombs, and R. L. Davis, eds., *Decision Processes*, pp. 49–59. Wiley, New York, 1954.

[Minsky, 1986] M. Minsky. *The Society of Mind.* Simon and Schuster, New York, 1986.

[Minton, 1990] S. Minton. Quantitative results concerning the utility of explanation-based learning. *Artificial Intelligence*, 42(2-3):363–391, 1990.

[Nagel, 1979] T. Nagel. The fragmentation of value. In *Mortal Questions*, chapter 9. Cambridge University Press, Cambridge, 1979.

[Newell, 1982] A. Newell. The knowledge level. *Artificial Intelligence*, 18(1):87–127, 1982.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA, 1988.

[Raiffa, 1968] H. Raiffa. *Decision Analysis: Introductory Lectures on Choices Under Uncertainty.* Addison-Wesley, Reading, MA, 1968.

[Rawls, 1971] J. Rawls. *A Theory of Justice.* Harvard University Press, Cambridge, MA, 1971.

[Russell and Wefald, 1989] S. Russell and E. Wefald. Principles of metareasoning. *Proc. First Int. Conf. on Principles of Knowledge Representation and Reasoning*, pp. 400–411, 1989.

[Savage, 1972] L. J. Savage. *The Foundations of Statistics.* Dover Publications, New York, 2nd ed., 1972.

[Schelling, 1980] T. C. Schelling. The intimate contest for self-command. *The Public Interest*, 60(Summer):94–118, 1980.

[Shoham, 1988] Y. Shoham. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence.* MIT Press, Cambridge, MA, 1988.

[Simon, 1955] H. A. Simon. A behavioral model of rational choice. *Quarterly J. of Economics*, 69:99–118, 1955.

[Simon, 1976] H. A. Simon. From substantive to procedural rationality. In S. J. Latsis, editor, *Method and Appraisal in Economics*, pp. 129–148. Cambridge University Press, 1976.

[Smith, 1986] D. E. Smith. Controlling inference. Report STAN-CS-86-1107, Department of Computer Science, Stanford University, 1986.

[Thaler and Shefrin, 1981] R. H. Thaler and H. M. Shefrin. An economic theory of self-control. *J. Political Economy*, 89(2):392–406, 1981.

[Thomason, 1986] R. H. Thomason. The context-sensitivity of belief and desire. *Reasoning about Actions and Plans: Proc. 1986 Workshop*, pp. 341–360, 1986.

[Touretzky *et al.*, 1987] D. S. Touretzky, J. F. Horty, and R. H. Thomason. A clash of intuitions: The current state of nonmonotonic multiple inheritance systems. *Proc. IJCAI-87*, pp. 476–482, 1987.

[Valiant, 1984] L. G. Valiant. A theory of the learnable. *CACM*, 18(11):1134–1142, 1984.

[Van Frassen, 1973] Bas C. Van Frassen. Values and the heart's command. *J. Philosophy*, LXX(1):5–19, 1973.

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

[Wellman *et al.*, 1989] M. P. Wellman, M. H. Eckman, C. Fleming, S. L. Marshall, F. A. Sonnenberg, and S. G. Pauker. Automated critiquing of medical decision trees. *Medical Decision Making*, 9:272–284, 1989.