Reset reproduction of article published in *Computational Intelligence*, Vol. 8, No. 2 (May 1992), pp. 376-409. Written May 2, 1990, revised January 24, 1992. Reprinted July 1994.

Rationality and its Roles in Reasoning

Jon Doyle DOYLE@LCS.MIT.EDU
Massachusetts Institute of Technology, Laboratory for Computer Science
Cambridge, Massachusetts 02139, U.S.A.

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.

1 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.

Judgments of human rationality commonly involve several different conceptions of rationality, including 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 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]

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 (notable early exceptions include Good's (1962) work on chess, Sproull's (1977) and Sproull and Feldman's (1977) work on strategy selection, and Gorry's (1968) work on diagnosis). 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.

[1] Of course, gambling may be rational if, for example, the gambler also seeks to satisfy non-monetary aims such as entertainment.

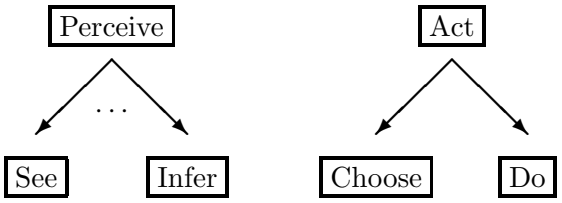

Figure 1: Intelligence involves both perception and action, which we take to involve both inference and choice.

The basic notions of economic rationality 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.

The purpose of this paper is to introduce the basic notions of economic rationality and to suggest that economic rationality should play as large a role as logical rationality in artificial intelligence. Though we support this claim later by pointing out many of the specific roles for rationality, the basic argument may be briefly summarized as follows. In classical terms, logic concerns Truth, while economics concerns Goodness (though a case can be made that neither says much about either), and judgments about both truth and goodness are crucial to intelligence. The argument also can be phrased in more modern terms. Intelligence involves both perception and action (see Figure 1). There are many types of perceptual actions (sight, touch, taste, etc.), one of which is inference, which we take to be a method for perceiving explicitly relationships that lie implicit in the agent's incomplete beliefs. Similarly, one may think of action as simply doing something. But in the usual way of viewing action, most actions are not determined by the agent's situation, but instead involve choices to do one thing rather than another. Thus both inference and choice are central operations in thinking.

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 remainder of the paper 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. These topics are studied in many fields besides AI (see, for example, (Bratman, 1987; Cherniak, 1986; Kahneman *et al.*, 1982)), and the immensity of the literature necessarily requires this brief survey to omit mention of many relevant works.

2 Economic rationality

The fundamental issue in the theory of economic rationality is *choice among alternatives*. Economic rationality simply means making "good" choices, where 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; Jeffrey, 1983; Savage, 1972).

2.1 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 $A \not\prec B$ to mean that $A \prec B$ does not hold and $A \precsim B$ (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. These preferences may change over time, but in this discussion we treat only instantaneous sets of preferences.

Rational agents choose *maximally preferred* alternatives. If $\mathcal{A} = \{A_1, A_2, \ldots\}$ 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 $A_j \in \mathcal{A}$. There may be several rational choices, or none at all if the set of preferences is inconsistent or if the set of alternatives is empty or infinite.

The theory does not require that a rational agent explicitly calculate or compute the maximality of its choices, only that the agent chooses alternatives that are in fact maximal according to its preferences. Just as stopped clocks are right twice a day, a choice made by a fixed rule which gives the same answer no matter what the circumstances will be rational nonetheless if that choice happens to be maximally preferred in the specific circumstances in question. For example, if one is traveling with a like-minded but more knowledgeable companion, it may be perfectly rational to imitate whatever the companion does without thinking much about it on the assumption that the companion is acting rationally (when in Rome, do as Romans do). We may thus distinguish such *implicitly* rational choices from *explicitly* rational choices made by explicitly calculating and comparing alternatives. In applying the theory of economic rationality in artificial intelligence, we normally seek to avoid explicitly rational reasoning procedures in favor of implicitly rational ones, in the expectation that the former require less computation than the latter.

The theory requires, as a minimum basis for rationality, that strict preference is a strict partial order (transitive and asymmetric), indifference is an equivalence relation (reflexive, symmetric, and transitive), and any two alternatives are either indifferent or one is preferred to the other, but not both at the same time. These separate requirements on preference and indifference may be captured more succinctly by requiring that weak preference be a complete preorder, or put 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.

2.2 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. The same set of preferences may be represented by many utility functions, as any strictly increasing transformation of a utility function will provide the same choices under maximization.

The economic notion of utility has its origins in Bentham's (1823) utilitarianism, but has since been divorced from his conceptions. Bentham based his theory on a hedonistic psychology in which actions were judged by the pleasure and pain they produced. According to Bentham, utility is just the difference between these effects, that is, the amount of pleasure minus the amount of pain. But this psychological conception of utility proved to be unworkable, as explaining everyday actions in terms of maximizing utility required hypothesizing some rather exotic pleasures and pains. Rather than abandon the basic approach, modern economics instead abandoned the psychological conception of utility, reaching its first modern formulation in the work of Pareto (1971).[2]

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. Computer science and AI often compare alternative computational methods according to their costs (usually time costs) alone. But the cost of some alternative says nothing about the value or benefits received from it, and neither cost nor benefit need be identical with the utility of the alternative if the risks posed by the action diminish (or increase) its attractiveness to the agent. Instead, the utility of an action is usually some function of the cost, benefit, risk, and other properties of the action.[3] 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, though it may independently affect the overall utility of the answer.

[2] Pareto developed the pure theory of preferences and showed how to construct utility functions corresponding to sets of preferences, given stringent assumptions about the preferences. Modern economics constructs markets and price systems from the equilibria determined by the utility functions of the individuals operating in the market. Mathematical economists have pursued seemingly endless work on how to construct equilibria under weaker assumptions about preferences. Some of this work may eventually prove useful in AI, which works with much less structured sets of preferences than those required in ordinary economic theory.

[3] The exception to the rule that costs are not utilities is the *opportunity cost* of a choice, which is defined to be the extra utility foregone (if any) by choosing the selected alternative rather than another. If the selected alternative is of maximal utility, its opportunity cost is zero, since no more valuable opportunities are passed up to take it.

In addition to distinguishing between cost and utility, the economic theory of utility also distinguishes the preferences of one agent from the preferences of other agents, a distinction sometimes confused in intuitive judgments of rationality according to the preferences of an observer (what we might call *external* utilities) rather than according to the preferences of the observed (what we might call *internal* utilities). For example, everyday judgments of rationality apparently go beyond the narrow, agent-relative criteria of logical and economic rationality when they criticize masochists or suicides for choosing to injure themselves on the basis of preferences the critic finds irrational.

More to the point for artificial intelligence, we must distinguish between the preferences of an artificial agent and the possibly different utilities the system has to its user (to whom patience, friendliness, reasonability, and convenience will be important), to its designer (to whom its longevity, salability, and maintainability will be important), to its informants, and to bystanders. Unlike natural agents which we view as simply having preferences, the preferences of artificial agents may be the result of design rather than a matter of interpretation. In many cases, designers construct agents presupposing agreement between their own preferences and those of the agent, but in other cases the agent preferences need not be the same as those of the designer. The two sets of preferences may differ due to designer error or a meaning-distorting implementation; they may differ by intent, as when one designer constructs competing agents; or they may differ because the artificial agent modifies its preferences based on its experiences.

One cannot define the notions of preference and utility purely in terms of beliefs and goals, for these are independent notions. As they are typically used in AI, goals only state what is desired, and do not give any information about the relative merits of different desirable alternatives (though one can of course assume that anything desired is to be preferred to anything not desired). But the objectives of most realistic tasks involve efficiency notions in crucial ways that render absolute notions of goal satisfaction inappropriate. For these tasks, the notions of preference and utility provide the necessary analytical concepts. On the other hand, finite sets of goals provide a focus for reasoning and action in a way that specifying utilities for the infinity of conceivable circumstances does not. To gain the advantages of both goals and preferences, one must introduce some new notion to connect the two. See (Wellman and Doyle, 1991) for an approach using multi-attribute descriptions of outcomes to define goals as conditions preferred to their opposites, holding other things equal.

2.3 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.[4] Compared with the basic theory, decision theory adds probability measures p_A which indicate the likelihood of each possible outcome for each alternative $A \in \mathcal{A}$. That is, decision theory supposes that the agent does not know the actual situation, but does have beliefs or expectations about the consequences of choices in different states. Decision theory also strengthens the notion of utility from an *ordinal* utility function u to a *cardinal*

[4]This theory has its origins in the work of Daniel Bernoulli (1738), and was mainly developed in this century by Ramsey (1931), de Finetti (1937), and Savage (1972).

utility function U, which imputes different values to each possible outcome. 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,

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

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 assumptions of decision theory can be formulated qualitatively. For example, Ramsey (1931) showed how to derive probability measures and preference orderings from the basic preference ordering $\precsim$ by interpreting the results of $\precsim$ applied to certain choices. He starts by defining a proposition to be "ethically neutral" if and only if any two worlds differing only in regard to its truth are of equal value, and (roughly speaking) characterizes probability $\frac{1}{2}$ by the degree of belief in an ethically neutral proposition such that the agent is indifferent to choosing alternatives of equal value conditional on the truth of the proposition. He goes on to define sets of alternatives which measure the probability of other propositions and reveal preferences among outcomes (see also (Savage, 1972; Jeffrey, 1983)). Alternatively, one might begin with an ordering $\precsim_p$ of possible outcomes according to their relative likelihood or degree of belief, and an ordering $\precsim_U$ of possible worlds according to their relative desirability and impose consistency and completeness conditions on their combination. In either approach, however, the result is a unique probability measure, and a class of cardinal utility measures closed under positive linear transformations. That is, cardinal utility measures are unique only up to choice of origin and unit, in that for any positive a and any b, the measure $aU + b$ yields the same behavior under maximization of expected utility as does U.

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.

3 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. Both uses are very important to artificial intelligence. Descriptive theories become increasingly important as the field matures and a greater emphasis is placed on gaining a thorough understanding of the nature and power of extant techniques. Normative theories, which were relatively less important in the early days of the field when the emphasis was on capturing the superficial phenomena and developing elementary techniques, become more important as the field seeks to extend its initial ideas, for only a normative theory can provide goals for deeper investigations.

To date, descriptive uses of logic have overshadowed descriptive uses of economic rationality. 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)). It also permits comparison of AI theories with formal psychological theories.

Normatively construed, however, logical and economic rationality provide different conceptions of how one should think, and logic has also overshadowed economic rationality as a normative theory of thinking. By their very nature, normative theories are more controversial than descriptive theories. To see more clearly the need for economic rationality as a normative theory in artificial intelligence, we begin by examining the normative use of logic, which we may call *logicism*.[5]

3.1 Logicism

The logicist approach to artificial intelligence views reasoning as a form of logical inference and seeks to construct 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

[5] Philosophers use the term "logicism" to mean something different, especially in the philosophy of mathematics where it names the thesis that mathematics is reducible to logic.

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. A purely logical reasoner guided only by the logicist rule would 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, and the reasoner needs to distinguish worthwhile inferences from worthless ones. 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, 1991)). 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.

3.2 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 among alternatives as in the logicist rule above. 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.[6] This apparent trivialization of his theory seems to be necessary: the theory is supposed to accurately characterize most of AI, and most AI systems ignore issues of rational choice in just the way Newell's principle suggests.

[6] "*Knowledge:* Whatever can be ascribed to an agent, such that its behavior can be computed according to the principle of rationality." (Newell, 1982, p. 105)

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. In the most general conception, 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!

Similarly, the more limited notion of *heuristic adequacy* (McCarthy and Hayes, 1969) advanced in the logicist approach involves using heuristics to determine the most useful sound inferences. But the notion of usefulness motivating the use of heuristics has rarely been made formal, which has sometimes brought work on heuristic methods into disrepute among logicians, mathematicians, and formally-minded AI theorists. In consequence, those who have insisted on formal theories have lacked any respectably formal alternative to logicism and sometimes have presented at least the appearance of subscribing to logicism, whether they actually do or not, while those who reject logicism sometimes also think it necessary to reject formality.

3.3 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 a formal theory of utility 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 utility among those open to the agent. When specialized to the decision-theoretic conception by augmenting utility with probability, 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 since the axioms of probability require that certain beliefs be consistent just as in logicism. But economic rationality imposes no requirement that inferences be logically sound.[7] 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

[7] 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 by identifying logic's notions of proposition and inference with psychological notions of belief and inference, and by imposing logic's standards of consistency and soundness on psychological belief states and inferential actions.

developing formulations of problems; economics plays a normative role in choosing both the problems to solve and the means of solving them.

4 Rationality in limited agents

The normative use of decision theory provides a standard for rationality, but one which (just like logical rationality) is often unattainable in practice because the theory is much too strong. It demands probabilities for all contingencies and preferences among all possibilities, but for many realistic problems, most of these probabilities and preferences are not available, and even if available are so numerous as to make their representation very difficult, if not impossible. In some cases, the available probabilities and preferences conflict with one another, and so cannot satisfy the consistency requirements of the theory. Moreover, calculating what alternatives are best can take a lot of time, so that the time for decision may have passed when the calculation is completed.[8]

In this section we explore some major limitations on agents that influence their degree of rationality. These include, in addition to limitations on the information and resources available for reasoning like those just mentioned, limitations due to the architecture or organization of the reasoner, physical limitations, and even some metaphysical limitations. While physical and metaphysical limitations are beyond human control, instantaneous informational, resource, and organizational limitations may be avoidable to some degree (these constraints can all be relaxed at a price), and the notion of *procedural* rationality arises through an attempt to use rational decisions about the control of reasoning to shape variable limitations in a way that maximizes the overall degree of rationality.

4.1 Informational limitations

The information available to a reasoner may be incomplete, inconsistent, indeterminate, or difficult to change. The reasoner's information may also be wrong, and while this certainly can lead to poor performance, incorrectness does not itself affect the rationality (logical or economic) of the agent. (Indeed, many people find misinformation a less embarrassing explanation for their mistakes than irrationality, which they interpret as stupidity.)

4.1.1 Incompleteness

Incompleteness of factual information about the world (including introspective information about oneself) is part of the human condition. The theory of rational choice explicitly assumes that some factual knowledge will be incomplete by incorporating probability theory to compare degrees of likelihood for unknown contingencies. At the same time, however, the theory of rational choice assumes the reasoner possesses complete and consistent information about the set of possible alternatives and, by postulating complete orderings of relative likelihood and desirability, about its own probabilities and preferences. These assumptions are much too strong in practice.

[8]The computational burdens imposed by economic rationality are not necessarily greater than those imposed by logical rationality, though both may take a lot of time when they are possible.

In the first place, the theory of rational choice requires that the agent be logically omniscient in the sense that that it believes all the consequences of its beliefs. In everyday life, however, people will assert or assent to some beliefs even though they profess ignorance about (or even deny) some of the logical consequences of the asserted beliefs. This seems to mean that people do not possess logical omniscience, and various attempts have been made toward formal theories of belief that do not demand that beliefs be closed under entailment (see (Levesque, 1984; Fagin and Halpern, 1985; Konolige, 1986)). Levesque (1984) introduced the term *explicit* belief to mean the beliefs to which an agent will assent, and the term *implicit* belief to mean the logical consequences of explicit belief. Thus ordinarily, the explicit beliefs of the agent will be incomplete.

Secondly, the implicit beliefs of the reasoner may be incomplete as well. That is, the reasoner may sometimes not know whether one circumstance is more likely than another. Similarly, it may not know which circumstance it prefers. Incompleteness of implicit beliefs and preferences means fundamental incompleteness, not simply incompleteness due to difficulty in determining consequences. This possibility is in fundamental conflict with the usual assumption of decision theory that each mental state determines complete and consistent sets of beliefs and preferences. Proponents of decision theory defend its assumptions by showing how one may take advantage of agents with incomplete beliefs and preferences by betting against them in ways that guarantee winning (so-called "Dutch book" arguments). But even if incompleteness is irrational (and the Bayesian arguments to this effect have many weak points), it certainly seems common in human reasoning. When people realize they are in situations that they have never considered before, they do not judge themselves to be irrational. Instead, they simply try to decide what beliefs and preferences to adopt (if any).

The incompleteness of the likelihood and preference orderings $\precsim_p$ and $\precsim_U$ means that these orders may be compatible with many probability and utility functions. We might rephrase the definition of rational choice to cover this situation by saying that a rational choice is one that is best no matter which compatible probability and utility functions are used. This works as long as the orders are complete with respect to the alternatives of interest, but if the orders fail to relate two alternatives of interest to each other, there need not be any rational choice since different completions will prefer each of these to the other. We might, however, call a choice *reasonable* in such circumstances if it is best according to some compatible probability and utility measures. Some authors (e.g., (Jaynes, 1979; Cheeseman, 1983)) suggest the further alternative of using a maximum entropy principle to fill in gaps in probability distributions.

4.1.2 Inertia

Standard treatments of decision theory assume that the beliefs of the agent persist over time in the absence of new information and that beliefs change according to the rule of Bayesian conditionalization (or some related rule (Harper, 1976)). But perfect sensitivity to new information is usually not possible because changing beliefs takes time and effort. If new information (produced either by perception, reasoning, or simple changes of mind) arrives too quickly, the reasoner may not be able to keep up, and will take actions that do not reflect all of the information it possesses. In such cases, the agent will appear to exhibit mental

inertia, persistence of beliefs in spite of changes, which can be interpreted as an insensitivity or sluggish response to new information. Even if persistence or nondegradation of beliefs without new information is good, persistence in spite of change may not be so good.[9] Thus a reasoner might be organized to anticipate the effort and risk involved in updating its beliefs by changing or maintaining them in the most rational manner possible. These rational forms of inertia include *conservatism* and *habit*.

Conservative reasoners retain as many beliefs as possible when new information leads to changes. For example, if the reasoner believes both P and Q, it could consistently accommodate a new belief $\neg(P \wedge Q)$ by giving up both P and Q, but this would give up more beliefs than necessary. A conservative revision might instead retain one of P or Q as beliefs. Conservatism is an important element of philosophical theories of belief revision and the related topic of counterfactuals (see (Quine and Ullian, 1978; Harman, 1986; Stalnaker, 1984; Gärdenfors, 1988)), and also appears as one of the central functions of a reason maintenance system (RMS, *née* truth maintenance system) (Doyle, 1979; Doyle, 1983b; Doyle, 1983a).

Habits may be viewed as implicitly rational procedures for making certain limited decisions. For example, a field-soldier's habit of ducking or diving to ground upon hearing the whistle of a shell in combat usually costs little, while taking the time to decide whether the shell really poses a danger sometimes exposes the soldier to deadly shrapnel. Habits seems to be just as important in understanding human behavior as is overt rational choice (see (Schumpeter, 1934; Nelson and Winter, 1982; Olson, 1982; Shils, 1981)). Habitual action can be quite complex and effective, as shown by Tinbergen (1951) and by recent work on reactive or situated systems (Agre and Chapman, 1987; Brooks, 1986; Rosenschein and Kaelbling, 1986), in which combinations of purely habitual rules are tailored to achieve the desired effects in a variety of circumstances.

4.1.3 Inconsistency

The theory of ideal rationality follows the logicist approach and requires that the beliefs and preferences of the reasoner be consistent. In practice, however, there are a number of reasons why the reasoner's beliefs or preferences about some subject may be inconsistent. Conflicting information may be gathered from multiple authorities and conflicting default rules may be contained in distinct but overlapping commonsense theories, and inconsistency may result if the reasoner has no indication of which authority or theory to trust on specific questions. But even if the reasoner could rid itself of inconsistency, it may not be rational to do so if detecting or removing the inconsistency would take too long (or otherwise not be worth the effort).

4.1.4 Indeterminacy

As noted previously, people do not always seem to have the complete sets of beliefs and preferences presumed by decision theory. But even when someone seems to exhibit beliefs

[9] Of course, the presence of noise requires modifying the notion of perfect sensitivity to mean perfect sensitivity to what the agent judges to be true information, not perfect sensitivity to what the agent judges to be noise. Some degree of resistance to change may in such cases serve to smooth the response of the agent, providing an implicitly rational means for displaying the appropriate degree of sensitivity.

or preferences concerning something, these attitudes do not always seem well-defined, in that one may observe behavior indicating that the relevant beliefs and preferences seem to vary with the agent's situation. Thomason (1986) calls this the *context sensitivity* of belief and desire. Context sensitivity then becomes overt indeterminacy if we take complex agents to inhabit several decision situations or contexts at the same time. While one might just think of such indeterminacy as simple inconsistency (believing both a proposition and its negation, or strictly preferring two things to each other), a richer theory seems in order, since at any instant the agent seems to believe just one thing, not two. Another possibility is that while the agent inhabits a single mental state, there are different possible ways of reading beliefs and preferences out of this state (perhaps one for each different context) so that the Ramsey-Savage constructions from revealed preferences need not give sensible results.

Though most of artificial intelligence presumes determinate beliefs and preferences, it does provide some theories exhibiting what one might think of as indeterminate belief states. The most notable examples are those theories of nonmonotonic reasoning which admit a multiplicity of possible states of belief corresponding to a single set of premises. For example, in default logic (Reiter, 1980), the knowledge of the reasoner is expressed in a set of axioms and a set of default rules. The beliefs of the agent must then be contained in the "extensions" of these axioms and rules, where the extensions are essentially sets of beliefs closed with respect to default rules and ordinary inference and generated from the initial axioms. Ordinarily, conflicts among default rules lead to multiple extensions from which one might choose the beliefs of the reasoner.[10] The problem of indeterminacy also arises in theories in which knowledge is distributed among different frames, perspectives, or subagents (see (Halpern, 1986; Minsky, 1975; Minsky, 1986)). As in the theory of distributed computer systems, there is usually no unique way to define the global states of such systems, either because the subsystems contain conflicting information or because there is no notion of simultaneous measurement across all subsystems. The standard theory of rationality would seem to require, however, that all these subsystems be mutually consistent and coordinated in their beliefs, in order that a consistent set of global beliefs may be defined. But since the point of distributing knowledge and reasoning is to free the reasoner from having to take all possibly relevant knowledge into account at every step, it is hard to imagine how (apart from using the shared environment as a coordinating store (Rosenschein and Kaelbling, 1986), which may not always be possible or successful) to ensure any degree of global consistency and coordination without undercutting this motivation for distributed computation.

4.2 Resource limitations

The most obvious limitations on resources facing reasoners are limitations on the time and memory available.

Time is limited in almost all realistic tasks, but there is usually no fixed upper limit to the time available. Different tasks will have different deadlines in different situations (for example, emergency room medical diagnosis has much shorter deadlines than diagnosis in the doctor's office), and deadlines can sometimes be postponed. More generally, the

[10] In (Doyle and Wellman, 1991), however, these multiple extensions are viewed as stemming from inconsistent preferences about belief.

reasoner may be able to trade other resources at its disposal for more time (for example, processor speeds might be changed if increased error rates are acceptable).

Computational reasoners also have finite memories, and most of AI assumes at the outset that all knowledge must be represented in finite structures or mental states. This causes difficulties for application of decision theory because the axioms of decision theory imply that the reasoner must possess the probabilities and utilities of infinitely many possibilities. To cut the problem of making rational decisions down to manageable size, Savage (1972) proposed the technique of only making decisions in what he called "small worlds," partitions of world states into equivalence classes corresponding to the relevant information. In sufficiently small worlds, simple qualitative orders may suffice to make the decisions, which helps bring decision theory into line with artificial intelligence practice.

Time and space need not be the only computational resources of interest. For example, properties of the input information such as degree of completeness or consistency may also be treated as resources which may be traded against time and space consumption. But more importantly, reasoners may have to take noncomputational resources into account in making decisions, in particular the effort required of humans or other cooperating agents, since a slow representation and inference system can tax the patience of a knowledge engineer, expert informant, or end user. For example, a cooperative agent may need to decide whether taking the time to draw some inference would annoy its user more than asking the user to answer another question.

4.3 Organizational limitations

In addition to resources, agents may be limited by their overall organization or architecture. For example, an architecture with twice the processing speed and memory size of another has the potential to be more rational. The underlying speed and size are only the crudest measures, however, and realistic assessments of relative rationality must look to the details of the architecture. Thus an architecture which permits reasoning to use lemmas and derived inference rules may be more rational than one that requires all inferences to start from the same base of general axioms and inference rules. Ignoring inputs and outputs, each architecture provides sets of states and actions, and these may be further analyzed into the structures or representations which may be encoded or realized in the states, and the operations, atomic and compound, which may be performed on these structures. The structures and operations determine the costs and reliabilities of different sorts of reasoning and representation tasks, so that the same resource allocations in one architecture will yield better performance than in another. In this way, we may view the very architecture itself as a limitation on the agent's capabilities.

One of the most important limitations imposed by an architecture is its degree of flexibility or adaptability. Some architectures, such as arithmetic logic units, provide fixed sets of possible structures and operations. But most architectures of interest in artificial intelligence permit representations and operations to be added to the repertoire or be modified. Virtually all architectures for reasoning permit addition of compound operations by being told (programmed), and many provide for learning new procedures and constructing plans as well. Many architectures also provide for compilation, in which new atomic and compound operations are created to adapt existing operations to special purposes and so increase their

efficiency for the intended or expected uses. Similarly, some architectures permit dynamic changes in the sorts of representations allowed. These changes may involve expansions of the set of possible representations, as when new symbols are created, but may just restrict the permissible representations by forbidding certain structures or combinations of structures. For example, database integrity constraints, which state conditions that the set of database entries must satisfy, may be viewed in this way (see (Reiter, 1988)), and the same holds for the reasons or justifications of reason maintenance systems, which represent conditions on what beliefs must be or must not be believed at the same time as others (see (Doyle, 1983b; Doyle, 1988a)).

More generally, we may call the legal operations and states of the agent its *constitution*, and divide the constitution into a fixed part unchangeable by the agent and a variable part under its control (see (Doyle, 1988a)). The fixed part often consists of a background logic, possibly very weak, that describes some of the consistency and closure properties of states, while the variable part can be viewed as a set of "laws of thought" or policies that rule out some states or changes which otherwise would be permitted by the background logic. For example, one may view integrity constraints and reason maintenance justifications as particular laws of thoughts carrying specific interpretations. N. Minsky (1988) has developed the notion of law into a general purpose programming framework, in which some of the organization of a program is explicitly described by a "law of the system."

4.4 Physical limitations

No list of limitations on reasoning abilities could be complete without mention of physical limitations. The fundamental physical limitations appear to be that information cannot be communicated faster than the speed of light, that quantities cannot be measured arbitrarily precisely at the same time, that the universe contains a finite amount of matter and energy in which to represent information, and that all matter and energy possesses inertia which restricts the rapidity of change. Most of these seem far removed from the more immediate limitations due to computational and informational resources, but this remoteness may not last long. The speed of light already warps some computer designs, and memory chip designers do not have far to go before they reach the quantum-theoretic limits on reliable information storage and retrieval. It is not inconceivable that these limits will someday directly influence some sorts of cognitive processing as well. Indeed, Penrose (1989) speculates that thinking depends on quantum gravitational effects.

4.5 Metaphysical limitations

In the preceding, we have made the usual assumption that rationality is a well-defined target. Economics certainly provides a precise formal conception of rationality, but the conception captured in this way may not, when all is said and done, seem to be the right idealization of ordinary intuitive conceptions of rationality. This may happen if there are several different intuitions underlying the notion of rationality. Milnor (1954), for example, listed a number of intuitively desirable properties of rational decisions under uncertainty, and then proved the set of these properties to be inconsistent. In particular, each standard approach satisfies some desirable properties but not others. If our intuitions about what rationality is are inconsistent, accurately capturing them in a precise theory will be difficult.

In the following, we consider the adequacy of the intuitions underlying decision theory. We first consider the strong evidence that expected utility does not capture a realistic notion of preference, either because expectations are nonlinear in the probabilities, or because utility is not cardinal, and then consider the possibility that overall utility functions simply do not exist.

Imperfections in the standard theories do not diminish the motivations stated previously for using economic rationality as a tool for understanding reasoning. Indeed, the concrete models of thinking developed in AI may prove instrumental in developing better theories of rationality. It is hard to tell why humans come to the decisions they do, since we know so little about the mind. But if the human conception of rationality is influenced by informational, resource, or organizational limitations, we may well expect computational intelligences to support similar judgments of rationality. Thus developing precise, concrete, and analyzable computational mechanisms that seem to behave in ways we consider rational may provide the insight needed to formulate an intuitively correct theory of rationality.

4.5.1 Is expected utility the right idealization?

The notion of economic rationality is supposed to idealize human behavior. Since humans fall short of the idealization in many ways, it is natural to ask if economic rationality is the right idealization of human behavior, or whether some of these apparent shortfalls are essential to what we mean by rational and should therefore be captured in a more appropriate idealization. Answering this question requires care for two reasons. The first reason is that, as Friedman (1953) has emphasized, the theory of economic rationality is a theory of results, not of process, so we cannot criticize the idealization because humans do not perform explicitly rational calculations. Thus divergences of some new ideal from economic rationality must be justified either by ideal considerations, or by arguing that some human limitations are so essential that they must be reflected in the ideal theory. Producing such arguments is difficult because of the second reason, namely that the basic theory of economic choice is so flexible that essentially any behavior can be rationalized. That is, it appears that for any history of behaviors one can always find a set of ordinal preferences so that every action of the history is rational (see (Stigler and Becker, 1977)). In the present setting, this means that the ideal of logical rationality may be completely undercut in some economically rational reasoners, since one may always be able to find sets of preferences which make rational any set or change of beliefs (cf. (Doyle, 1991a)).

In spite of these difficulties, there does seem to be reason to doubt that economic rationality is the best idealization of human behavior. For example, psychological experiments indicate that humans do not average utilities in the way called for by decision theory's equation (1), which says that expected utility depends linearly and additively on the probabilities of events. Other discrepancies include so-called preference reversals, framing, and anchoring effects (see (Gärdenfors and Sahlin, 1988; Kahneman *et al.*, 1982; Machina, 1987; Machina, 1989)). Many theorists find it difficult to view these discrepancies as human failings since even very intelligent humans (e.g., famous economists who would seem to have a stake in the matter) persist in disputing the prescriptions of decision theory, even when fully informed and aware of all the arguments. But if these experts make the decisions we

think they *should* make, then we must have in mind a normative theory of rationality that differs from the normative interpretation of economic rationality.

4.5.2 Pluralism

The basic theory of rational choice supposes that preferences may be described by a utility function. We criticized this assumption above because it is not always practical to ensure that preferences are complete. But this incompleteness may be more than a practical matter, for the existence of a utility function supposes that all values can be combined into a single scale of desirability, and this may not always be possible.

One of the oldest questions taken up by philosophers is whether there are many things, or just one, and in particular, whether there are many ultimate goods or some overarching single good. Plato, for example, believed in a single value system covering everything, but some philosophers argue that there are several incomparable final goods, which cannot be combined into a single scale except by idiosyncratic invention (see (Nagel, 1979; Van Fraassen, 1973)).

The usual approach to choice when there are several different ways of comparing things is multiattribute decision theory, in which utilities are functions of vectors of attribute valuations (see (Keeney and Raiffa, 1976)). But multiattribute decision theory does not prescribe how to combine the different attributes into a utility measure. In fact, it appears that there can be no general uncontroversial combination method, as an analogue of Arrow's social choice (group decision making) impossibility theorem holds for the problem of combining attribute valuations (Arrow, 1963; Arrow and Raynaud, 1986). According to this result, the only "reasonable" combination procedures that always yield rational results are "dictatorial" procedures in which one attribute's valuation always determines the whole utility. Thus the decision analyst must consider the tradeoffs among the different attributes and decide on a combination method. This means that, if the pluralists are right, the construction of utility functions depends on the whims of individuals. The practical consequence for AI of these considerations is that there may not be any system of inference good for all tasks in all applications (see (Horty *et al.*, 1990; Doyle and Wellman, 1991)).

4.6 Coping with limitations

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. Consider, for example, an agent which has two decision-making methods available: flipping a coin and explicitly calculating the rational choice. If the agent has unlimited calculating abilities (or no deadlines to meet), these methods will be equally costly, and flipping coins will be irrational since that method will produce a rational choice only with probability less than 1, while explicit calculation may produce a rational choice with probability 1. If the agent's abilities or time is limited, coin-flipping may be much cheaper than explicit calculation, and the utility of these different methods will depend on the circumstances. If

explicit calculation takes more time than is available, and if both late and incorrect answers have minimal utility, then the random decision method will have higher expected utility than explicitly calculating a rational answer, so flipping coins will be rational.

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 (as in (Doyle, 1980)).[11] Thus striking a balance between control and reasoning computations means taking effort expended at all these levels into account. In practice, the deliberative information available at higher levels but unavailable at lower ones vanishes 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. Jeffrey (1983) calls such decisions *ratified* decisions, as the beliefs and preferences that result from the choice themselves ratify the choice as rational. Rawls (1971) uses the term *reflective equilibrium* to describe a similar notion, in which the agent's beliefs and principles are in agreement (that is, it has just the set of beliefs and principles it thinks it should have).

5 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,

[11] Since all information available at one level is potentially available at all higher levels, one need not organized deliberation into towers of reflective deliberations. Lipman (1989) shows that these towers of decisions are equivalent to base-level decisions (which may be enormous or even uncomputable).

	God exists	doesn't
Believe	$+\infty$	$-\epsilon$
Doubt	$-\infty$	$+\epsilon$

Figure 2: Pascal's utility assessments of the possible consequences of his decision about belief in God. The entries $+\infty$ and $-\infty$ represent the infinite utility and disutility, respectively, of eternal salvation and damnation, while ϵ represents the finite amount of pleasure enjoyed or foregone due to belief during his life.

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?[12]

We here enumerate some tasks in which rational choice would seem to play a significant role. Substantial work has already been done on some of these, but others have seen only initial explorations. We emphasize again that we expect automatic procedures for these tasks to involve explicit calculations of rational choices only when we cannot find implicitly rational procedures of comparable or greater expected utility to the agent or to the designer.

5.1 Rational assumptions and belief revision

Decisions about what to believe in cases of uncertainty may be interpreted as implicitly rational assessments of the expected utility of different states of belief. The classic example of a rational decision to believe is that of Pascal's "wager" (1962), in which Pascal decides to believe in God because he expects the utility of believing to exceed (infinitely) the utility of disbelieving (see Figure 2). As William James argued concerning the "will to believe" (James, 1897), similar considerations can be offered in justification of many mundane cases of belief. Note that in making his decision, Pascal clearly assessed the utility of accurate and erroneous belief and doubt about God's existence, not the utility of God's existence. The everyday pitfall of wishful thinking occurs when one chooses a belief by comparing the utilities of a condition and its opposite (e.g., "I believe I'll pass the test because it'll be terrible if I don't") rather than the utilities of believing and doubting the condition (e.g., "I mustn't assume I'll pass the test because otherwise I'll slack off and fail it").

[12] Hansson and Mayer (personal communication) claim preliminary results along these lines; see also Russell and Wefald's (1991) interpretation of conspiracy numbers.

In artificial intelligence, decisions to believe commonly take the form of default rules for making assumptions. Several authors have suggested justifying these rules in decision-theoretic terms (Doyle, 1983b; Doyle, 1989a; Langlotz and Shortliffe, 1989; Shoham, 1988) as "compiled" rational decisions about reasoning given expectations and preferences about the inferences to be performed over some interval or while performing some task. Ginsberg (1991) has gone beyond these suggestions to provide formal examples of such justifications by analyzing some simple formal models of how the use of default rules can reduce the cost of planning. Specifically, he estimates the cost of planning with and without defaults that provide plausible planning "islands" (intermediate states that must later be justified by additional, more detailed plans), and derives analytical relationships indicating cases in which using defaults reduces the overall cost of planning.

In spite of these economic justifications of default rules, there has been little direct work on the task of mechanizing the formulation and adoption of rational default rules, which is necessary to make "compiling" of rational decisions an activity of the agent itself rather than an activity of its designer. This task involves choosing the appropriate context for reasoning, and in particular, choosing the appropriate statistical reference classes (that is, the relevant populations and subpopulations) (Loui, 1986; Loui, 1988), since different default rules are appropriate for different subpopulations. Put this way, however, this task has much in common with the problems of learning useful concepts and finding useful taxonomic organizations, offering the possibility of adapting techniques developed to address those problems.

Rationally choosing assumptions is closely related to the more general problem of rationally updating or revising beliefs as new information is gained. The usual approach, as mentioned earlier, is to make belief revisions conservative, minimizing the changes made in the set of beliefs (Harman, 1986; Gärdenfors, 1988). There is no logical justification for conservatism as a principle of belief revision. For example, the frame problem is a problem precisely because logic alone says nothing about persistence of belief. Indeed, the mere truth of some belief is no reason for holding it in memory, for otherwise memory could easily be exhausted recording tautologies. Instead, economics, not logic, provides the proper justification for conservatism. The reason for recording a belief in memory is that it is valuable, so beliefs should not be removed from memory unless the reasoner expects them to undermine the efficacy of actions. For example, the reasoner might retain some belief in 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.[13] Alternatively, the reasoner may expect that removing the belief might produce ambiguity that increases the costs of reasoning by necessitating search. The reasoner might even keep useless beliefs in memory if the effort involved in identifying them outweighs the value of the storage consumed. Far from being an irrational resistance to rational change, the inertia represented by conservatism is a perfectly rational response to the costs of updating beliefs.

At present we lack a precise economic theory of conservative belief revision. Formal theories of conservative belief revision typically adopt principles like minimizing the number

[13]Memory costs can extend beyond storage space alone if storing items in memory slows down inference in retrieval, if fixed memory size introduces opportunity costs for storing specific information, or if memory decay forces the reasoner to refresh its beliefs periodically.

or set of changed beliefs (as in (Harman, 1986)), or axiomatize the logical relationships that must hold between prior beliefs, new information, and posterior beliefs (as in (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. Accordingly, it would be valuable to develop theories of belief revision in which the agent rationally chooses among alternative revisions (see, for example, (Doyle, 1991a; Doyle and Wellman, 1990)).

Rational choice would be especially valuable in the special case of backtracking, or choosing which assumptions to abandon in order to restore consistency to a set of beliefs. Backtracking can be viewed as searching through the consistent subsets of a body of information, or through all subsets if determining consistency is difficult. In the first place, most work in artificial intelligence attempts to maintain consistent beliefs, often without regard for the cost involved in doing so, and a more careful assessment of this approach is needed. In some cases, we may expect that analyzing and removing the inconsistency will not be worth the effort it requires. In the second place, the most "efficient" backtracking schemes, such as dependency-directed backtracking (Stallman and Sussman, 1977; Doyle, 1979), are those based on conservative belief revision, in which consistency restored by discarding as few old (or new) beliefs as necessary to produce a consistent set. These methods are "efficient" only in comparison with nonconservative schemes like chronological backtracking, which may discard many beliefs that are unrelated to the inconsistency. But even dependency directed backtracking is irrational in that it chooses revisions arbitrarily from the possible alternatives without checking to see if one logically possible alternative is preferred to another (Doyle, 1983b; Doyle, 1988a). Some backtracking techniques do use restricted sorts of ordering information, but none have been designed to choose revisions rationally on the basis of the reasoner's preferences.

One factor that may complicate development of an adequate treatment of belief revision is that the preferences guiding revisions must themselves be revised. For example, a reasoner may initially consider two assumptions equally desirable, but if a lot of evidence is found for one but none is found for the other, the reasoner may eventually prefer the assumption with the stronger evidence. In addition, some preferences might be based on assumptions, and so may need to be changed if the underlying assumptions change. Present theories of belief revision, however, say little about how changes of belief can or should change preferences guiding belief revision. For example, Gärdenfors and Makinson (1988) show how to characterize belief revisions by ordering beliefs according to state-dependent degrees of "epistemic entrenchment," where the least entrenched beliefs are abandoned first when necessary, but do not explain how entrenchment orderings should vary with the state of belief (cf. (Doyle, 1991b)).

5.2 Rational representations of inconsistent information

Backtracking, whether rational or irrational, aims at keeping the set of beliefs consistent by changing beliefs. An alternative approach to dealing with inconsistent information is to *represent* (in the sense of (Doyle, 1988a; Doyle, 1989a)) the inconsistent information with one of its consistent subsets, that is to keep the full body of information and decide how to

reason or act by first choosing a consistent subset of the information and then using this consistent information to choose how to reason or act.

The main application of this technique so far has been in theories of inheritance and nonmonotonic logics. These give the appearance of consistency by means of *credulous* representations, in which maximal consistent subsets represent inconsistent sets of rules, and *skeptical* representations, in which the intersection of all maximal consistent subsets represents the inconsistent information (Horty *et al.*, 1990). The skeptical approach amounts to reasoning without the conflicting information, while the credulous approach amounts to temporarily taking sides in the conflict. Skeptical approaches are preferred by some authors because they avoid the risk of taking the wrong side in the conflict. But in so doing they assume the risk of not taking the right side in the conflict. Neither skepticism nor credulity is rational in all situations, so one topic for investigation is how to choose a representation appropriate to the reasoner's situation (an approach which makes beliefs context-sensitive in the sense of Thomason (1986) discussed earlier.) This problem appears to have close connections with the problem of group decision making, and appears to suffer from the same metaphysical difficulties (Doyle and Wellman, 1991).

Practical approaches based on credulous representations must also involve a notion of conservatism. If the consistent representation is chosen anew for each action, different actions may be based on inconsistent subsets even if the underlying beliefs have not changed. This would mean that the inconsistency of the agent's beliefs would be apparent in the inconsistency of the agent's actions. (Freedom from this embarrassment is one of the attractions of the skeptical approach.) Alternatively, the consistent representation may be updated conservatively as new information is obtained, so that actions are based on the consistent representation as long as it is still consistent with new information. This approach gives the appearance of consistency.

The use of representations of inconsistent sets of beliefs has close relations with Levesque's (1984) notion of explicit belief, and more specifically, with the notions of *manifest* and *constructive* belief (Doyle, 1989b) which divide explicit beliefs into the manifest or explicitly represented beliefs and the constructive or implicitly represented beliefs derived when needed. The derivation or construction of constructive belief involves rational choice, but a fully satisfactory formal definition of rational constructive belief remains to be developed, as the self-referential nature of the choice requires use of equilibrium notions like ratified decisions (Jeffrey, 1983).

5.3 Rational approximations

Imprecise knowledge is an important special case of ignorance, since reasoners often lack the details of answers to questions even when they possess the essentials of the answers. Since the degree of precision of a solution is sometimes very important, a number of approaches have been developed to improve imprecise answers. Rather than simply trying to find the exact answer, these approaches attempt to provide answers of increasing degrees of precision or worth by incremental means.

The most familiar notion of approximation is that of monotone approximation of complete answers, which means simply that the reasoner constructs increasingly more complete partial solutions, so that each successive partial solution extends all previous ones. This

notion plays a central role in the modern theory of data structures and recursive computation (see (Scott, 1982)). It appears in work on deduction and planning under the names constraint-based reasoning and constraint posting (see, for example, (Stallman and Sussman, 1977; Saraswat, 1989; Stefik, 1981)). But the completeness of answers may be only one factor in their utility, so that a more complete partial answer may no better than (or even worse than) a less complete one. Accordingly, recent research has developed approximation methods in which the aim is that utility or probability of partial answers increase monotonically as further effort is applied.

Monotone approximation of answers of maximal utility means constructing partial solutions incrementally so that the utility of the partial solutions (not necessarily their information content) grows monotonically. Horvitz (1988) calls such processes *flexible computations*, while Dean and Boddy (1988) call them *anytime algorithms* because they can be interrupted anytime to get as good an answer as possible in the time available. Russell and Wefald (1991) achieve similar results by always having a default action available. Computation toward a better action may continue, but the default action is always available to be taken, and is replaced only when a better substitute is found.

Monotone approximation of answers of maximal probability means constructing partial solutions so that the probability of correctness of the partial solutions grows monotonically. This is almost the inverse of approximating the most informative answer, since a vague answer is more probable than any more precise answer refining it. Monotone probability approximations have been pursued actively in recent work on learning, specifically on so-called *probably approximately correct* (PAC) algorithms that produce approximately correct answers with high probability (see (Valiant, 1984; Haussler, 1988)). These algorithms combine the earlier notions of probabilistic algorithms, including Monte Carlo algorithms (which quickly produce an answer that is usually right) and Las Vegas algorithms (which produce a correct answer, usually quickly). PAC algorithms produce definitions of concepts from samples so that the definitions produced are, with a specified degree of certainty, in agreement with the sampled concepts to a specified percentage overall. If one measures value by degree of correctness, one can view PAC definitions as rational approximations to concepts. One challenge for AI is to generalize this notion, if possible, to notions like PAC inferences from axioms, PAC plans for achieving goals, or even probably approximately maximal choices (which might be taken as a version of Simon's notion of satisficing).

5.4 Rational search and inference

The problems of reasoning and search with limited time have received substantial attention from AI researchers. One version of this problem is that of real-time reasoning and search, in which a fixed amount of time is available for each episode of reasoning or search (Breese and Fehling, 1990; Korf, 1988). Other work has explored obtaining predictable response time, rather than fixed response times (Levesque and Brachman, 1987). We will here highlight only some recent work on using techniques from decision theory and operations research in allocating effort within time limitations, which may be soft deadlines in addition to hard deadlines. See (Dean and Wellman, 1991, Chapter 8) for a more detailed survey, and (Good, 1962; Simon and Kadane, 1975) for some early discussions of rational control of search.

Russell and Wefald (1991) present their approach to rationally controlling A* and other search methods as a general theory of "metareasoning." In their approach, the system maintains a default action (e.g., a chess move), which is just the action that currently appears to be best in terms of carrying the system towards winning the game. The system must choose what to do from among a set of computational actions. These might have the effect of causing the system to change its default move to a better one, or might leave the default action unchanged. Since all computations consume time, the aim is to take only those computational actions which change the default to a better action, or more specifically, those actions which result in the greatest estimated increase in value. It is not possible to tell with certainty what the result of a computation is without performing it, since if it were, perfect chess could be played without making any computations at all. Thus the system must choose based on its expectations about the estimated changes in utility due to different computational actions. Russell and Wefald develop explicit formulas for these quantities, and proceed to show how these general formulas may be applied in practice by making several simplifying assumptions. These include representing the computational utility of an action as the difference between the "intrinsic" utility of the action and a measure of time cost; assuming the only available computational actions are search steps (the "meta-greedy" assumption) and estimating their ultimate effect by assuming at most one more search step will be taken (the "single-step" assumption); and assuming that individual computational actions change the utility estimate for at most one alternative.

Each of the simplifying assumptions made by Russell and Wefald represents a significant restriction on the generality of their technique. For example, Hansson and Mayer (1989) report that the single-step assumption can cause the utility of the search to degenerate to worse than that of random search. They suggest improving the quality of utility estimates by using the values revealed by the position evaluation function as evidence for the true values, and by treating the search graph as a belief network so as to combine this evidence via Bayesian conditionalization. In addition, one may seek to control reasoning on a larger scale than individual search steps using similar decision-theoretic judgments (see (Breese and Fehling, 1990; Fehling *et al.*, 1989; Horvitz, 1988; Horvitz *et al.*, 1989)).

The procedure that Russell and Wefald employ for assessing the value of further deliberation works quickly by employing efficient numerical encodings of the results of statistical experiments with searches from chess positions. But for other sorts of reasoning, one cannot assume that this metalevel decision can be made as easily. For example, in logical inference, it can be very difficult to estimate how frequently different inference goals occur, how frequently different methods succeed, and how much time these methods take. Smith (1986; 1988) has analyzed a number of important special cases of this problem, but much work remains to be done. But even when one does know how long inference operations take, allocating effort can be difficult, as is shown by Etzioni's (1991) work on rational control of traditional goal-directed search. Etzioni presents an efficient method based on sorting for constructing plans for single goals, but shows that constructing rational plans for multiple goals is NP-complete. This necessitates the use of approximate optimization algorithms, of which there are many that perform acceptably in practice.

The main source of difficulty in allocating effort to work on multiple tasks is that some computations are monolithic, that is, they produce no useful results until they are completely finished. Even worse, some computations must be allowed to finish if the knowledge

base is to be returned to a consistent state. These problems become much simpler if computations can be interrupted (and perhaps restarted) at any time and still produce useful partial or approximate answers while leaving the knowledge base in an acceptable state. This approach forms the basis of the previously mentioned methods for monotone approximation of choices of maximal utility (Horvitz, 1988; Dean and Boddy, 1988).

5.5 Rational learning

Learning involves selection as well as memorization, but most theories of learning pursued in AI pay little attention to these choices. For example, while DeJong's (1983) original criteria for explanation-based learning concerned utility of the result, many techniques of explanation-based generalization (Mitchell *et al.*, 1986) use only a categorical notion of operationality of concept definitions that avoids comparing the different explanations that might form the basis of learning. Similarly, many techniques have been developed for case-based reasoning, but few if any of these have been analyzed or interpreted in terms of the expected utility of the cases recorded or of the reconstruction process. Indeed, many reasoning systems record aspects of past episodes of reasoning indiscriminately (using, for example, reason maintenance systems or chunking (Laird *et al.*, 1987)), and this leads to clogging memory with records of dubious value (Minton, 1990). More generally, few learning theories address the problem of how to choose when or whether to learn, and so fail to cover the deliberate learning so common in everyday life.

The subject of learning seems ripe for reexamination with an eye to analyzing the rationality of different methods for deciding what to memorize, what to forget, what information to summarize, what summaries to memorize, how carefully or precisely to classify or categorize new objects in taxonomies, and how to most efficiently organize taxonomies. In addition to examining the rationality of traditional methods for these tasks, the new techniques of PAC learning need scrutiny as well. As mentioned earlier, identifying degree of correctness with utility permits one to view PAC learning as rational selection of concept definitions, but it is not apparent that PAC definitions are the same as definitions of maximal expected utility, even under this identification.

5.6 Rational planning

Planning involves several tasks (in addition to reasoning and search) in which rational choice may play a significant role (cf. (Doyle and Wellman, 1990)). As noted earlier, the very notion of "goal" is suspect, as goals often implicitly involve efficiency considerations. Thus one of the main roles for rational choice is comparison of alternative plans (see (Feldman and Sproull, 1977; Smith, 1988; Langlotz, 1989)). For example, Wellman (1990) describes a planner which uses *dominance* relations among sets of plans to guide the search and to isolate the fundamental tradeoffs among methods. Dominance relations among plan classes are derived from preferences among individual plans. For instance, according to Savage's (1972) "sure-thing" principle, one plan should be preferred to another if its outcome is preferred to the outcome of the other in every possible state of nature, that is, if it is of greater expected utility no matter what probability distribution is used. (Wellman's system employs a qualitative generalization of this principle.) The point is that dominated plans or classes of plans may thus be pruned from the search tree, as they will never be the best

choice. Using dominance relations can simplify the search for plans substantially, since it permits separating the planning process into two stages: first determining the underlying tradeoffs among methods that will remain valid even if the details of the situation change, and then filling out the details of the plan once one has chosen how to resolve the tradeoffs involved.

Expected utility also provides a criterion for deciding whether to plan for contingencies. One cannot reasonably plan for all contingencies, since there are usually far too many, most of which are remote possibilities. A better approach is to plan for contingencies only when the expected utility of preparing the plan exceeds not planning for it. This means assessing the probability of different contingencies, and constructing plans only for those contingencies for which the costs of planning ahead (mainly opportunity costs) are less important than the costs of deciding what to do at the time the contingency occurs.

5.7 Rational self-organization

It would be valuable to develop methods for rationally choosing what patterns of reasoning and search to compile into new operations and representations, or constitutional logics and laws of thought. For example, Breese and Horvitz (1990) analyze the tasks of reformulating and compiling probabilistic networks in decision-theoretic terms, and Kant's (1983) system for choosing data-structures uses both frequency and cost information, but set within a matrix of heuristic decision methods that might be improved by use of decision-theoretic techniques. In general, one can expect methods for choosing new representations and operations should be closely related to those developed for choosing how to summarize memories, how to choose default rules, and how to plan for contingencies.

6 Other tasks

A number of more general problems must be addressed to enable rapid progress on the specific applications of rational choice discussed above.

6.1 Automate decision formulation and analysis

Decision theory and economics do not address the problem of how to formulate the alternatives, probabilities, and preferences appropriate to different tasks. Decision formulations lie outside the scope of these theories, and must be supplied by the person using the theory. The field of *decision analysis* (Howard and Matheson, 1984; Raiffa, 1968) has developed methodologies for eliciting, estimating, and evaluating decision formulations from human experts. Not coincidentally, much of this work bears striking similarity to the more recent efforts in AI on developing expert systems, as both fields address essentially the same problem with somewhat different tools. Recent work has evidenced a convergence of method, with decision analysts adopting AI tools, and expert systems employing decision-analytic concepts (see (Horvitz *et al.*, 1988)). While decision-analytic methodologies offer a starting point, autonomous reasoners cannot always rely on obtaining their decision formulations through external decision analysts. 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).

6.2 Improve representations for probabilities and preferences

Even when the reasoner needs only finitely many beliefs, there may still be too many probabilities to store as explicit representations. In medical diagnosis problems, for example, the number of conditional probabilities needed in the worst case is exponential in the number of diseases and symptoms or tests results (Szolovits and Pauker, 1978).

This difficulty was taken for some time to indicate the impracticality of probabilistic representations of knowledge, but several methods have been developed recently for succinctly specifying large amounts of probabilistic information. These representations simply describe some of the constraints relating different classes of probabilities, and rely on specific inference techniques to recover the rest, usually against a background of independence assumptions that implicitly determine most possible conditional probabilities (Horvitz *et al.*, 1988; Pearl, 1988)). Unfortunately, it can sometimes be hard to carry out inferences in these frameworks, even though they greatly restrict what can be expressed (see (Cooper, 1990)). Wellman (1990) goes even further in this direction and provides more general qualitative descriptions of probability distributions, including qualitative probabilistic influences and qualitative synergies. Extending and improving these succinct probabilistic representations constitutes an important open problem.

The problem of succinctly representing preferences and utility information is similar in nature and importance to that of finding good representations for probabilistic information, but has received much less attention. Some initial qualitative representations have been proposed in (Wellman and Doyle, 1991; Doyle *et al.*, 1991), which exploit multi-attribute descriptions of outcomes to define notions of preference *ceteris paribus*, that is, preference holding all else equal. As with representing probabilistic information, much more work is needed here.

6.3 Identify realistic cost measures

While the standard complexity classes (e.g., P and NP) support an elegant and rich mathematical theory, they are usually inappropriate measures of cost for the purposes of judging rationality. In the first place, virtually all economic constraints on resources are much tighter than those characterizing the usual complexity classes. The basic time available for some task is often some fixed amount, and can be increased beyond that constant only at great cost. For example, the class of all chess strategies computable with at most 8 ply lookahead, or that evaluate at most 10 million positions, may be of much greater practical interest than P to designers of chess players. Secondly, the usual complexity classes contain the same functions no matter what machine is of interest, even though the same problems may have very different costs on different machines. Finally, the most-cited complexity classes are those which measure worst case costs, and the worst case is often not of great practical interest compared with the average or normal cases. For example, apparently NP-hard problems are solved every day in industry with no special trouble, in some cases

because the worst case is very rare to begin with, and in others because practitioners follow tacit conventions that usually prevent the worst cases from arising.

One of the important open problems is identification of realistic and appropriate measures of costs for reasoning and representation, specifically, measures that depend on the architecture involved.[14] In practice, each different computational or cognitive architecture provides different primitives, and a pattern of reasoning may require many more fundamental operations in one architecture than in another. Moreover, the actual time required by each of these may also vary widely even if the same number of steps is required in each architecture. To be useful in designing and controlling reasoners, the cost measures used should include measures of cognitive effort, expressed in the different sorts of operations provided by the architecture, as well as the usual notions of time and space. This means that different measures may be needed for different architectures.

6.4 Investigate choice without utility

If there really are numerous independent basic values guiding choices of an agent, or if the agent has a distributed mental organization, global utility functions may not exist and overall rationality may be impossible to define. Even so, we may seek to identify weaker notions of rationality or reasonability appropriate to these circumstances. That is, we might attempt to define rationality in a way that does not presume utility functions.

One set of candidate idealizations can be obtained by viewing the fragmented values from the social choice perspective and asking what behaviors result as we abandon or relax some of the rationality principles for combining different dimensions of value. For example, we could simply give up on global rationality, permitting the overall preferences of the reasoner to be intransitive or inconsistent. Intransitive preferences mean that the reasoner may keep changing its mind about which alternative seems best, and in some cases this sort of inconsistency may allow outsiders to take advantage of the reasoner. Direct inconsistency may lead to the same problems, or to the agent not being able to find any maximally preferred alternatives at all, depending on how it is treated. The difficulties caused by such irrationality might be minimal if the reasoner always appears to be locally rational. By this we mean that the agent appears to have consistent beliefs and preferences in each context or episode of reasoning, even if different beliefs and preferences are apparent in different contexts or episodes. Another approach is to impose a ranking on the independent values. In general, we may try to estimate the likelihoods of different sorts of choices arising, and estimate the risks that different sorts of irrationality entail in each of these cases.

6.5 Exploit economic theory

Economics has substantial theories of common ways of organizing human societies or businesses, such as bureaucracies, markets, and more general social decision-making frameworks mixing authority relationships with decisions among equals. In many cases, there are direct analogies between these organizational designs and common organizations for representation and reasoning systems. For example, subroutine hierarchies correspond to bureaucracies,

[14] A related project might be to reconstruct the theory of computation using classes defined by utilities as opposed to costs.

and use of an intelligent "big switch" or homunculus corresponds to central planning. AI needs to exploit these theories.

Economic theory provides formal evidence for various comparative advantages and disadvantages among the different possible organizations that are sometimes assumed to hold, usually without any sort of rigorous argument, among the different computational organizations. For example, bureaucracies achieve efficiency at certain tasks at the cost of overall inflexibility. Sometimes this inflexibility makes these systems easier to analyze and predict, but often it merely serves to make them brittle or unreliable. This brittleness is especially apparent in the computational setting, since when one procedure hands control to a subordinate, it becomes a slave to the subordinate: if the subordinate does not return, the calling procedure cannot recover and proceed. To avoid this pitfall of unthinking bureaucratic behavior, subprocedures must be implemented as tools or assistants that can be ignored or discarded when necessary rather than as subroutines (as in (Minsky, 1986)). Similarly, markets are highly regarded by many economists since they offer some ways of overcoming the limitations of knowledge of individual agents, while central planning is often deplored, as it presumes the existence of an all-knowing, perfectly rational human who makes all the complicated decisions the populace at large is presumably incompetent to make on its own. But market organizations have received only initial exploration in AI (Huberman, 1988; Wellman, 1992).

Not all of the economists' theories will be directly applicable to the case of mental organization, and some of them may not be applicable at all, but it seems imprudent to ignore the economists completely since they have spent many years gaining knowledge, performing analyses, and developing mathematical tools for the study of rational organizations. In particular, some economists have turned their attention to cognitive issues by applying economic models to modeling the attitudes and mental organization of the individual agent (see, for example, (Becker, 1976; Elster, 1979; Hirschman, 1982; Kydland and Prescott, 1977; Schelling, 1984a; Schelling, 1984b; Thaler and Shefrin, 1981)), and these would seem especially important for study.

The most relevant general topics are microeconomics, which treats the theory of markets composed of individual rational agents (see (Henderson and Quandt, 1980; Debreu, 1959)) and social choice theory, which treats nonmarket or social decision-making schemes such as voting (see (Mueller, 1989)). At this point, macroeconomics, which treats certain questions of collective behavior and group psychology, may be less directly relevant, except possibly to connectionists. (Macroeconomics is, however, what most people think of when they think of economics, as it is what usually fills the newspaper accounts of economic and political debates.)

6.6 Design cooperative architectures

Many traditional AI architectures are essentially programming systems offering a fixed functionality. In order to achieve the desired performance, the user must tailor the axiomatization and the phrasing of goals, keeping the system's functionality in mind. Often these systems are advertised as "general purpose" systems, but this description is misleading. True generality means that the system can be *adapted* into a tool useful for any specific purpose. This adaptability involves a great degree of flexibility. But the "generality" in de-

signer's minds is often just the opposite, namely a fixed functionality or level of performance chosen to be on average reasonably useful no matter which purpose is intended, that is, a functionality independent of the differing needs of different users in different situations. For example, some knowledge representation systems are specialized for the worst case, which makes for substantial inconvenience in normal cases. The fact is that there is no universally good fixed functionality, as any fixed functionality will be a Procrustean bed, good for some purposes but bad for others. True generality would seem to come only by having the system take the user's changing purposes into account and choosing how to tailor or manage its services so that they maximize utility to the user. That is, to be maximally useful, systems should "put themselves in the user's place" by adopting the user's expectations and preferences and making decisions rationally with these adopted attitudes (Doyle and Patil, 1991).

6.7 Design provably rational architectures

AI has developed numerous general architectures for representation and reasoning (such as SOAR (Laird *et al.*, 1987) or PRODIGY (Minton *et al.*, 1989)), but we lack at present an understanding of whether any of these architectures are rational in any precise sense. A careful analysis may reveal that none is very rational in the economic sense, but even if some existing architectures are rational to some degree, demonstrating that may be very difficult since they were not designed with an eye toward theoretical analysis. It may be that architectures based on explicit decision-theoretic comparisons (such as SUDO-PLANNER (Wellman, 1990)) or on other operations of demonstrated rationality may be easier to analyze. Accordingly, much might be learned by attempting to design architectures that provide the same ranges of possible behaviors but are structured so as to permit clean theoretical analyses of their fundamental properties.

6.8 Find useful measures of degrees of rationality

Since there are many different dimensions along which agents may be limited, and since one can sometimes trade improved performance along one dimension for worsened performance along another, defining a precise notion of degree of rationality for comparing agents may be difficult. An easier task is to compare limitations along each dimension separately, for example, according to the completeness of beliefs and preferences separately, and also according to the strength of underlying inferential systems (Doyle, 1988a). It would also be interesting to define a notion of relative rationality, in analogy with the notion of relative computability. This might form the basis of a theory of rationality that clearly identifies the mechanizable degrees of rationality.

6.9 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 of lectures in an introductory AI class is probably not adequate.

7 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 the capabilities and behaviors of heuristic systems in terms as precise as those of logic, and to do this there is no alternative apparent to the formal tools of the economic theory of rational choice. 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. Thinking is costly, and to gain the efficiency in thought that intelligence seems to require, it seems necessary for agents to engage in some form of what one might call rational self-government (Doyle, 1988a). As an added benefit, using the theory of rational choice to recast artificial intelligence ideas makes it easier to relate these ideas to relevant concepts in philosophy, psychology, economics, decision theory, statistics, operations research, management, political theory, sociology, and anthropology. Rationality is a central notion in each of these fields, and by using the same language they do, artificial intelligence may find more thorough developments of some of its informal ideas, may find some new ideas to apply to formalizing thinking, and perhaps can see more easily which of its own ideas can make a contribution to these other fields (Doyle, 1988b).

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. This means finding ways of guiding reasoning that are rational to some degree, where the means of guidance is itself only rational to some degree. 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

This paper is an extended version of an invited talk (Doyle, 1990) presented at AAAI-90. I thank Ramesh Patil, Peter Szolovits, and Michael Wellman for reading drafts, Rich Thomason for lending me some of his notes, Tom Dean, Othar Hansson, Eric Horvitz, Barton Lipman, Andrew Mayer, Stuart Russell, Joseph Schatz, and David Smith for valuable discussions, and the referees for helpful suggestions. This work was supported by the National Library of Medicine through National Institutes of Health Grant No. R01 LM04493.

References

Agre, P. E. and Chapman, D. 1987. Pengi: An implementation of a theory of activity. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pp. 268–272.

Arrow, K. J. 1963. *Social Choice and Individual Values.* Yale University Press, second edition.

Arrow, K. J. and Raynaud, H. 1986. *Social Choice and Multicriterion Decision-Making.* MIT Press, Cambridge, Massachusetts.

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

Becker, G. S. 1976. *The Economic Approach to Human Behavior.* University of Chicago Press, Chicago.

Bentham, J. 1823. *Principles of Morals and Legislation.* Oxford University Press, Oxford. Originally published in 1789.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii academiae scientiarum imperialis Petropolitanae* (for 1730 and 1731), 5:175–192.

Bratman, M. 1987. *Intention, Plans, and Practical Reason.* MIT Press, Cambridge, MA.

Breese, J. S. 1987. *Knowledge Representation and Inference in Intelligent Decision Systems.* Research Report 2, Rockwell International Science Center, Palo Alto, CA.

Breese, J. S. and Fehling, M. R. 1990. Control of problem solving: Principles and architecture. In Schacter, R. D., Levitt, T. S., et al., editors, *Uncertainty in Artificial Intelligence 4*, volume 9 of *Machine Intelligence and Pattern Recognition*, pp. 59–68. North-Holland, Amsterdam.

Breese, J. S. and Horvitz, E. J. 1990. Ideal reformulation of belief networks. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, pp. 64–72.

Brooks, R. A. 1986. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23.

Cheeseman, P. 1983. A method of computing generalized Bayesian probability values for expert systems. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp. 198–202.

Cherniak, C. 1986. *Minimal Rationality.* MIT Press, Cambridge, MA.

Cooper, G. F. 1990. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2-3):393–405.

de Finetti, B. 1937. La prévision: see lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7.

Dean, T. and Boddy, M. 1988. An analysis of time-dependent planning. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pp. 49–54.

Dean, T. L. and Wellman, M. P. 1991. *Planning and Control.* Morgan Kaufmann, San Mateo, CA.

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

DeJong, G. 1983. Acquiring schemata through understanding and generalizing plans. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp. 462–464.

Doyle, J. 1979. A truth maintenance system. *Artificial Intelligence*, 12(2):231–272.

Doyle, J. 1980. *A Model for Deliberation, Action, and Introspection.* AI-TR 581, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, Cambridge, MA.

Doyle, J. 1983a. The ins and outs of reason maintenance. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp. 349–351.

Doyle, J. 1983b. *Some Theories of Reasoned Assumptions: An Essay in Rational Psychology.* Technical Report 83-125, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA.

Doyle, J. 1988a. *Artificial Intelligence and Rational Self-Government.* Technical Report CS-88-124, Computer Science Department, Carnegie-Mellon University, Pittsburgh, PA.

Doyle, J. 1988b. Big problems for artificial intelligence. *AI Magazine*, 9(1):19–22.

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

Doyle, J. 1989b. Mental constitutions and limited rationality. In Fehling, M. and Russell, S., editors, *Papers of the AAAI Symposium on AI and Limited Rationality*, pp. 18–22.

Doyle, J. 1990. Rationality and its roles in reasoning (extended abstract). In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pp. 1093–1100.

Doyle, J. 1991a. Rational belief revision (preliminary report). In Fikes, R. E. and Sandewall, E., editors, *Proceedings of the Second Conference on Principles of Knowledge Representation and Reasoning*, pp. 163–174.

Doyle, J. 1991b. Reason maintenance and belief revision: Foundations vs. coherence theories. In Gärdenfors, P., editor, *Belief Revision.* Cambridge University Press, Cambridge. To appear.

Doyle, J. and Patil, R. S. 1991. Two theses of knowledge representation: Language restrictions, taxonomic classification, and the utility of representation services. *Artificial Intelligence*, 48(3):261–297.

Doyle, J., Shoham, Y., and Wellman, M. P. 1991. A logic of relative desire (preliminary report). In Ras, Z. W. and Zemankova, M., editors, *Methodologies for Intelligent Systems, 6*, volume 542 of *Lecture Notes in Artificial Intelligence*, pp. 16–31, Springer-Verlag, Berlin.

Doyle, J. and Wellman, M. P. 1990. Rational distributed reason maintenance for planning and replanning of large-scale activities. In Sycara, K., editor, *Proceedings of the DARPA Workshop on Planning and Scheduling*, pp. 28–36, Morgan Kaufmann, San Mateo, CA.

Doyle, J. and Wellman, M. P. 1991. Impediments to universal preference-based default theories. *Artificial Intelligence*, 49(1-3):97–128.

Elster, J. 1979. *Ulysses and the Sirens: Studies in Rationality and Irrationality.* Cambridge Univerisity Press, Cambridge.

Etzioni, O. 1991. Embedding decision-analytic control in a learning architecture. *Artificial Intelligence*, 49(1-3):129–159.

Fagin, R. and Halpern, J. Y. 1985. Belief, awareness, and limited reasoning: Preliminary report. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pp. 491–501.

Fehling, M. R., Einav, D., and Breese, J. S. 1989. Adaptive planning and search. In *Proceedings of the AAAI Symposium on AI and Limited Rationality*, pp. 46–54.

Feldman, J. A. and Sproull, R. F. 1977. Decision Theory and Artificial Intelligence II: The hungry monkey. *Cognitive Science*, 1:158–192.

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics.* University of Chicago Press.

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

Gärdenfors, P. and Makinson, D. 1988. Revisions of knowledge systems using epistemic entrenchment. In Vardi, M. Y., editor, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, pp. 83–95.

Gärdenfors, P. and Sahlin, N.-E., editors. 1988. *Decision, Probability, and Utility: Selected Readings.* Cambridge University Press, Cambridge.

Ginsberg, M. L. 1991. The computational value of nonmonotonic reasoning. In Allen, J., Fikes, R., and Sandewall, E., editors, *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pp. 262–268.

Good, I. J. 1962. A five-year plan for automatic chess. In Dale, E. and Michie, D., editors, *Machine Intelligence 2*, volume 2, pp. 89–118. Oliver and Boyd, London.

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

Gorry, G. A. and Barnett, G. O. 1968. Sequential diagnosis by computer. *Journal of the American Medical Association*, 205(12):849–854.

Halpern, J. Y. 1986. Reasoning about knowledge: An overview. In Halpern, J. Y., editor, *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the 1986 Conference*, pp. 1–17, Morgan Kaufmann, Los Altos, CA.

Hansson, O. and Mayer, A. 1989. Heuristic search as evidential reasoning. In *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence*, pp. 152–161.

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

Harper, W. L. 1976. Rational belief change, Popper functions, and counterfactuals. In Harper, W. L. and Hooker, C. A., editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, volume 1, pp. 73–115. Reidel, Dordrecht.

Haussler, D. 1988. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial Intelligence*, 36(2):177–221.

Henderson, J. M. and Quandt, R. E. 1980. *Microeconomic Theory: A Mathematical Approach.* McGraw-Hill, New York, third edition.

Hirschman, A. O. 1982. *Shifting Involvements: Private Interest and Public Action.* Princeton University Press, Princeton.

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

Horty, J. F., Thomason, R. H., and Touretzky, D. S. 1990. A skeptical theory of inheritance in nonmonotonic semantic networks. *Artificial Intelligence*, 42(2-3):311–348.

Horvitz, E. J. 1988. Reasoning under varying and uncertain resource constraints. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pp. 111–116.

Horvitz, E. J., Breese, J. S., and Henrion, M. 1988. Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning*, 2:247–302.

Horvitz, E. J., Cooper, G. F., and Heckerman, D. E. 1989. Reflection and action under scarce resources: Theoretical principles and empirical study. In Sridharan, N. S., editor, *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, volume 2, pp. 1121–1127, San Mateo, CA. International Joint Conferences on Artificial Intelligence, Inc., Morgan Kaufmann.

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

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

James, W. 1897. *The Will to Believe and Other Essays in Popular Philosophy.* Longmans, Green, and Co., New York.

Jaynes, E. T. 1979. Where do we stand on maximum entropy? In Levine and Tribus, editors, *The Maximum Entropy Formalism.* M.I.T. Press.

Jeffrey, R. C. 1983. *The Logic of Decision.* University of Chicago Press, Chicago, second edition.

Kahneman, D., Slovic, P., and Tversky, A., editors. 1982. *Judgment under Uncertainty: Heuristics and Biases.* Cambridge University Press.

Kant, E. 1983. On the efficient synthesis of efficient programs. *Artificial Intelligence*, 20(3):253–305.

Keeney, R. L. and Raiffa, H. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* John Wiley and Sons, New York.

Konolige, K. 1986. What awareness isn't: a sentential view of implicit and explicit belief. In Halpern, J. Y., editor, *Proceedings of the Conference on Theoretical Aspects of Reasoning about Knowledge*, pp. 241–250.

Korf, R. 1988. Real-time heuristic search: New results. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pp. 139–144.

Kydland, F. E. and Prescott, E. C. 1977. Rules rather than discretion: the inconsistency of optimal plans. *J. Political Economy*, 85(3):473–491.

Laird, J. E., Newell, A., and Rosenbloom, P. S. 1987. SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33:1–64.

Langlotz, C. P. 1989. *A Decision-Theoretic Approach to Heuristic Planning.* Report No. STAN-CS-89-1295, Computer Science Department, Stanford University, Stanford, CA.

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

Langlotz, C. P., Shortliffe, E. H., and Fagan, L. M. 1986. Using decision theory to justify heuristics. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 215–219.

Levesque, H. J. 1984. A logic of implicit and explicit belief. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 198–202.

Levesque, H. J. and Brachman, R. J. 1987. Expressiveness and tractability in knowledge representation and reasoning. *Computational Intelligence*, 3:78–93.

Lipman, B. 1989. How to decide how to decide how to ...: Limited rationality in decisions and games. In *Working Notes of the AAAI Symposium on AI and Limited Rationality*, pp. 77–80.

Loui, R. P. 1986. Interval-based decisions for reasoning systems. In Kanal, L. N. and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence*, pp. 459–472. North-Holland.

Loui, R. P. 1988. Computing reference classes. In Lemmer, J. F. and Kanal, L. N., editors, *Uncertainty in Artificial Intelligence 2*, pp. 273–289. North-Holland.

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

Machina, M. J. 1989. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, XXVII(4):1622–1668.

McCarthy, J. and Hayes, P. J. 1969. Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B. and Michie, D., editors, *Machine Intelligence 4*, pp. 463–502. Edinburgh University Press.

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

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

Minsky, M. 1975. A framework for representing knowledge. In Winston, P. H., editor, *The Psychology of Computer Vision*, chapter 6, pp. 211–277. McGraw-Hill, New York.

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

Minsky, N. H. 1988. *Law-Governed Systems.* Technical report, Rutgers University, Computer Science Department, New Brunswick, NJ.

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

Minton, S., Carbonell, J. G., et al. 1989. Explanation-based learning: A problem solving perspective. *Artificial Intelligence*, 40(1-3):63–118.

Mitchell, T. M., Keller, R. M., and Kedar-Cabelli, S. T. 1986. Explanation-based generalization: a unifying view. *Machine Learning*, 1(1):47–80.

Mueller, D. C. 1989. *Public Choice II.* Cambridge University Press, Cambridge, second edition.

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

Nelson, R. R. and Winter, S. G. 1982. *An Evolutionary Theory of Economic Change.* Harvard University Press, Cambridge, MA.

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

Olson, M. 1982. *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities.* Yale University Press, New Haven, CT.

Pareto, V. 1971. *Manual of Political Economy.* Kelley, New York. Originally published 1927. Translated by A. S. Schwier, edited by A. S. Schwier and A. N. Page.

Pascal, B. 1962. *Pensées sur la religion et sur quelques autres sujets.* Harvill, London. Translated by M. Turnell, originally published 1662.

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

Penrose, R. 1989. *The Emperor's New Mind: Concerning Computers, Minds, and The Laws of Physics.* Oxford University Press, New York.

Quine, W. V. and Ullian, J. S. 1978. *The Web of Belief.* Random House, New York, second edition.

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

Ramsey, F. P. 1931. Truth and probability. In Braithwaite, R. B., editor, *The Foundations of Mathematics and Other Logical Essays*, pp. 156–198. Routledge and Kegan Paul, London.

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

Reiter, R. 1980. A logic for default reasoning. *Artificial Intelligence*, 13:81–132.

Reiter, R. 1988. On integrity constraints. In Vardi, M. Y., editor, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, pp. 97–111.

Rosenschein, S. J. and Kaelbling, L. P. 1986. The synthesis of digital machines with provable epistemic properties. In Halpern, J. Y., editor, *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the 1986 Conference*, pp. 83–98.

Russell, S. and Wefald, E. 1991. Principles of metareasoning. *Artificial Intelligence*, 49(1-3):361–395,.

Saraswat, V. 1989. *Concurrent Constraint Programming Languages.* PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania.

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

Schelling, T. C. 1984a. The intimate contest for self-command. In *Choice and Consequence: Perspectives of an errant economist*, pp. 57–82. Harvard University Press, Cambridge.

Schelling, T. C. 1984b. The mind as a consuming organ. In *Choice and Consequence: Perspectives of an errant economist*, pp. 328–346. Harvard University Press, Cambridge.

Schumpeter, J. A. 1934. *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle.* Harvard University Press, Cambridge. Translated by R. Opie.

Scott, D. S. 1982. Domains for denotational semantics. In *Proceedings of the International Conference on Automata, Languages, and Programming.*

Shils, E. 1981. *Tradition.* Chicago University Press, Chicago.

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

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

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

Simon, H. A. and Kadane, J. B. 1975. Optimal problem-solving search: All-or-none solutions. *Artificial Intelligence*, 6:235–247.

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

Smith, D. E. 1988. *A Decision Theoretic Approach to the Control of Planning Search.* Technical Report LOGIC-87-11, Department of Computer Science, Stanford University.

Sproull, R. 1977. *Strategy Construction Using a Synthesis of Heuristic and Decision-Theoretic Methods.* PhD thesis, Stanford University.

Stallman, R. M. and Sussman, G. J. 1977. Forward reasoning and dependency-directed backtracking in a system for computer-aided circuit analysis. *Artificial Intelligence*, 9(2):135–196.

Stalnaker, R. C. 1984. *Inquiry.* MIT Press, Cambridge, MA.

Stefik, M. 1981. Planning with constraints (molgen: Part 1). *Artificial Intelligence*, 16:111–140.

Stigler, G. J. and Becker, G. S. 1977. De gustibus non est disputandum. *American Economic Review*, 67:76–90.

Szolovits, P. and Pauker, S. G. 1978. Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence*, 11:115–144.

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

Thomason, R. H. 1986. The context-sensitivity of belief and desire. In Georgeff, M. P. and Lansky, A. L., editors, *Reasoning about Actions and Plans: Proceedings of the 1986 Workshop*, pp. 341–360. Morgan Kaufmann.

Tinbergen, N. 1951. *The Study of Instinct.* Clarendon Press, Oxford.

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

Van Fraassen, B. C. 1973. Values and the heart's command. *Journal of Philosophy*, LXX(1):5–19.

Wellman, M. P. 1990. *Formulation of Tradeoffs in Planning Under Uncertainty.* Pitman and Morgan Kaufmann.

Wellman, M. P. 1992. A general equilibrium approach to distributed transportation planning. To appear.

Wellman, M. P. and Doyle, J. 1991. Preferential semantics for goals. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 698–703.

Wellman, M. P., Eckman, M. H., et al. 1989. Automated critiquing of medical decision trees. *Medical Decision Making*, 9(4):272–284.

Contents