1 Rationality

Rationality, according to some, is an excess of reasonableness. We should be rational enough to confront the problems of life, but there is no need to go whole hog. Indeed, doing so is something of a vice. Isaac Levi, *The Covenant of Reason* (Cambridge University Press, 1997)

The disciplines of science and engineering are complementary. Science comes from the Latin root *scientia*, or knowledge, and engineering comes from the Latin root *ingenerare*, which means to beget. While any one individual may fulfill multiple roles, a scientist *qua* seeker of knowledge is concerned with the analysis of observed natural phenomena, and an engineer *qua* creator of new entities is concerned with the synthesis of artificial phenomena. Scientists seek to develop models that explain past behavior and predict future behavior of the natural entities they observe. Engineers seek to develop models that characterize desired behavior for the artificial entities they construct. Science addresses the question of how things are; engineering addresses the question of how things might be.

Although of ancient origin, science as an organized academic discipline has a history spanning a few centuries. Engineering is also of ancient origin, but as an organized academic discipline the span of its history is more appropriately measured by a few decades. Science has refined its methods over the years to the point of great sophistication. It is not surprising that engineering has, to a large extent, appropriated and adapted for synthesis many of the principles and techniques originally developed to aid scientific analysis.

One concept that has guided the development of scientific theories is the "principle of least action," advanced by Maupertuis[1] as a means of systematizing Newtonian mechanics. This principle expresses the intuitively pleasing notion that nature acts in a way that gives the greatest effect with the least effort. It was championed by Euler, who said: "Since the fabric of the world is the most perfect and was established by the wisest Creator, nothing happens in this world in which some reason of maximum or minimum

[1] Beeson (1992) cites Maupertuis (1740) as Maupertuis' first steps toward the development of this principle.

would not come to light" (quoted in Polya (1954)).[2] This principle has been adopted by engineers with a fruitful vengeance. In particular, Wiener (1949) inaugurated a new era of estimation theory with his work on optimal filtering, and von Neumann and Morgenstern (1944) introduced a new structure for optimal multi-agent interactivity with their seminal work on game theory. Indeed, we might paraphrase Euler by saying: "Nothing should be designed or built in this world in which some reason of maximum or minimum would not come to light." To obtain credibility, it is almost mandatory that a design should display some instance of optimization, even if only approximately. Otherwise, it is likely to be dismissed as *ad hoc*.

However, analysis and synthesis are inverses. One seeks to take things apart, the other to put things together. One seeks to simplify, the other to complicate. As the demands for complexity of artificial phenomena increase, it is perhaps inevitable that principles and methods of synthesis will arise that are not attributable to an analysis heritage – in particular, to the principle of least action. This book proposes such a method. It is motivated by the desire to develop an approach to the synthesis of artificial multi-agent decision-making systems that is able to accommodate, in a seamless way, the interests of both individuals and groups.

Perhaps the most important (and most difficult) social attribute to imitate is that of coordinated behavior, whereby the members of a group of autonomous distributed machines coordinate their actions to accomplish tasks that pursue the goals of both the group and each of its members. It is important to appreciate that such coordination usually cannot be done without conflict, but conflict need not degenerate to competition, which can be destructive. Competition, however, is often a byproduct of optimization, whereby each participant in a multi-agent endeavor seeks to achieve the best outcome for itself, regardless of the consequences to other participants or to the community.

Relaxing the demand for optimization as an ideal may open avenues for collaboration and compromise when conflict arises by giving joint consideration to the interests of the group and the individuals that compose the group, provided they are willing to accept behavior that is "good enough." This relaxation, however, must not lead to reliance upon *ad hoc* rules of behavior, and it should not categorically exclude optimal behavior. To be useful for synthesis, an operational definition of what it means to be good enough must be provided, both conceptually and mathematically. The intent of this book is two-fold: (a) to offer a criterion for the synthesis of artificial decision-making systems that is designed, from its inception, to model both collective and individual interests; and (b) to provide a mathematical structure within which to develop and apply this criterion. Together, criterion and structure may provide the basis for an alternative view of the design and synthesis of artificial autonomous systems.

[2] Euler's argument actually begs the question by using superlatives (most perfect, wisest) to justify other superlatives (maximum, minimum).

1.1 Games machines play

Much research is being devoted to the design and implementation of artificial social systems. The envisioned applications of this technology include automated air-traffic control, automated highway control, automated shop floor management, computer network control, and so forth. In an environment of rapidly increasing computer power and greatly increased scientific knowledge of human cognition, it is inevitable that serious consideration will be given to designing artificial systems that function analogously to humans. Many researchers in this field concentrate on four major metaphors: (a) brain-like models (neural networks), (b) natural language models (fuzzy logic), (c) biological evolutionary models (genetic algorithms), and (d) cognition models (rule-based systems). The assumption is that by designing according to these metaphors, machines can be made at least to imitate, if not replicate, human behavior. Such systems are often claimed to be intelligent.

The word "intelligent" has been appropriated by many different groups and may mean anything from nonmetaphorical cognition (for example, strong AI) to advertising hype (for example, intelligent lawn mowers). Some of the definitions in use are quite complex, some are circular, and some are self-serving. But when all else fails, we may appeal to etymology, which owns the deed to the word; everyone else can only claim squatters rights. *Intelligent* comes from the Latin roots *inter* (between) + *legĕre* (to choose). Thus, it seems that an indispensable characteristic of intelligence in man or machine is an ability to choose between alternatives.

Classifying "intelligent" systems in terms of anthropomorphic metaphors categorizes mainly their syntactical, rather than their semantic, attributes. Such classifications deal primarily with the way knowledge is represented, rather than with the way decisions are made. Whether knowledge is represented by neural connection weights, fuzzy set-membership functions, genes, production rules, or differential equations, is a choice that must be made according to the context of the problem and the preferences of the system designer. The way knowledge is represented, however, does not dictate the rational basis for the way choices are made, and therefore has little to do with that indispensable attribute of intelligence.

A possible question, when designing a machine, is the issue of just where the actual choosing mechanism lies – with the designer, who must supply the machine with all of rules it is to follow, or with the machine itself, so that it possesses a degree of true autonomy (self-governance). This book does not address that question. Instead, it focuses primarily on the issue of *how* decisions might be made, rather than *who* ultimately bears the responsibility for making them. Its concern is with the issue of how to design artificial systems whose decision-making mechanisms are understandable to and viewed as reasonable by the people who interface with such systems. This concern leads directly to a study of rationality.

This book investigates rationality models that may be used by men or machines. A rational decision is one that conforms either to a set of general principles that govern preferences or to a set of rules that govern behavior. These principles or rules are then applied in a logical way to the situation of concern, resulting in actions which generate consequences that are deemed to be acceptable to the decision maker. No single notion of what is acceptable is sufficient for all situations, however, so there must be multiple concepts of rationality. This chapter first reviews some of the commonly accepted notions of rationality and describes some of the issues that arise with their implementation. This review is followed by a presentation of an alternative notion of rationality and arguments for its appropriateness and utility. This alternative is not presented, however, as a panacea for all situations. Rather, it is presented as a new formalism that has a place alongside other established notions of rationality. In particular, this approach to rational decision-making is applicable to multi-agent decision problems where cooperation is essential and competition may be destructive.

1.2 Conventional notions

The study of human decision making is the traditional bailiwick of philosophy, economics, and political science, and much of the discussion of this topic concentrates on defining what it means to have a degree of conviction sufficient to impel one to take action. Central to this traditional perspective is the concept of preference ordering.

Definition 1.1
Let the symbols "$\succeq$" and "$\cong$" denote binary ordering relationships meaning "is at least as good as" and "is equivalent to," respectively. A **total ordering** of a collection of options $U = \{u_1, \ldots, u_n\}$, $n \geq 3$, occurs if the following properties are satisfied:

Reflexivity: $\forall u_i \in U: u_i \succeq u_i$

Antisymmetry: $\forall u_i, u_j \in U: u_i \succeq u_j \;\&\; u_j \succeq u_i \Rightarrow u_i \cong u_j$

Transitivity: $\forall u_i, u_j, u_k \in U: u_i \succeq u_j,\ u_j \succeq u_k \Rightarrow u_i \succeq u_k$

Linearity: $\forall u_i, u_j \in U: u_i \succeq u_j$ or $u_j \succeq u_i$

If the linearity property does not hold, the set U is said to be **partially ordered**. □

Reflexivity means that every option is at least as good as itself, antisymmetry means that if u_i is at least as good as u_j and u_j is at least as good as u_i, then they are equivalent, transitivity means that if u_i is as least as good as u_j and u_j is at least as good as u_k, then u_i is at least as good as u_k, and linearity means that for every u_i and u_j pair, either u_i is at least as good as u_j or u_j is at least as good as u_i (or both).

1.2.1 Substantive rationality

Once in possession of a preference ordering, a rational decision maker must employ general principles that govern the way the orderings are to be used to formulate decision rules. No single notion of what is acceptable is appropriate for all situations, but perhaps the most well-known principle is the classical economics hypothesis of Bergson and Samuelson, which asserts that individual interests are fundamental; that is, that social welfare is a function of individual welfare (Bergson, 1938; Samuelson, 1948). This hypothesis leads to the doctrine of **rational choice**, which is that "each of the individual decision makers behaves as if he or she were solving a constrained maximization problem" (Hogarth and Reder, 1986b, p. 3). This paradigm is the basis of much of conventional decision theory that is used in economics, the social and behavioral sciences, and engineering. It is based upon two fundamental premises.

P-1 *Total ordering*: the decision maker is in possession of a total preference ordering for all of its possible choices under all conditions (in multi-agent settings, this includes knowledge of the total orderings of all other participants).

P-2 *The principle of individual rationality*: a decision maker should make the best possible decision for itself, that is, it should optimize with respect to its own total preference ordering (in multi-agent settings, this ordering may be influenced by the choices available to the other participants).

Definition 1.2
Decision makers who make choices according to the principle of individual rationality according to their own total preference ordering are said to be **substantively rational**. □

One of the most important accomplishments of classical decision theory is the establishment of conditions under which a total ordering of preferences can be quantified in terms of a mathematical function. It is well known that, given the proper technical properties (e.g., see Ferguson (1967)), there exists a real-valued function that agrees with the total ordering of a set of options.

Definition 1.3
A **utility** ϕ on a set of options U is a real-valued function such that, for all $u_i, u_j \in U$, $u_i \succeq u_j$ if, and only if, $\phi(u_i) \geq \phi(u_j)$. □

Through utility theory, the qualitative ordering of preferences is made equivalent to the quantitative ordering of the utility function. Since it may not be possible, due to uncertainty, to ensure that any given option obtains, orderings are usually taken

with respect to expected utility, that is, utility that has been averaged over all options according to the probability distribution that characterizes them; that is,

$$\pi(u) = E[\phi(u)] = \int_U \phi(u) P_C(du),$$

where $E[\cdot]$ denotes mathematical expectation and P_C is a probability measure characterizing the random behavior associated with the set U. Thus, an equivalent notion for substantive rationality (and the one that is usually used in practice) is to equate it with maximizing expected utility (Simon, 1986).

Not only is substantive rationality the acknowledged standard for calculus/probability-based knowledge representation and decision making, it is also the *de facto* standard for the alternative approaches based on anthropomorphic metaphors. When designing neural networks, algorithms are designed to calculate the *optimum* weights, fuzzy sets are defuzzified to a crisp set by choosing the element of the fuzzy set with the *highest* degree of set membership, genetic algorithms are designed under the principle of survival of the *fittest*, and rule-based systems are designed according to the principle that a decision maker will operate in its own *best* interest according to what it knows.

There is a big difference in perspective between the activity of analyzing the way rational decision makers make decisions and the activity of synthesizing actual artificial decision makers. It is one thing to postulate an explanatory story that justifies how decision makers might arrive at solution, even though the story is not an explicit part of the generative decision-making model and may be misleading. It is quite another thing to synthesize artificial decision makers that actually live such a story by enacting the decision-making logic that is postulated. Maximizing expectations tells us what we may expect when rational entities function, but it does not give us procedures for their operation. It may be instructive, but it is not constructive.

Nevertheless, substantive rationality serves as a convenient and useful paradigm for the synthesis of artificial decision makers. This paradigm loses some of its appeal, however, when dealing with decision-making societies. The major problem is that maximizing expectations is strictly an individual operation. Group rationality is not a logical consequence of individual rationality, and individual rationality does not easily accommodate group interests (Luce and Raiffa, 1957).

Exclusive self-interest fosters competition and exploitation, and engenders attitudes of distrust and cynicism. An exclusively self-interested decision maker would likely assume that the other decision makers also will act in selfish ways. Such a decision maker might therefore impute self-interested behavior to others that would be damaging to itself, and might respond defensively. While this may be appropriate in the presence of serious conflict, many decision scenarios involve situations where coordinative activity, even if it leads to increased vulnerability, may greatly enhance performance. Especially when designing artificial decision-making communities, individual rationality may not be an adequate principle with which to characterize desirable behavior in a group.

The need to define adequate frameworks in which to synthesize rational decision-making entities in both individual and social settings has led researchers to challenge the traditional models based on individual rationality. One major criticism is the claim that people do not usually conform to the strict doctrine of substantive rationality – they are not utility maximizers (Mansbridge, 1990a; Sober and Wilson, 1998; Bazerman, 1983; Bazerman and Neale, 1992; Rapoport and Orwant, 1962; Slote, 1989). It is not clear, in the presence of uncertainty, that the best possible thing to do is always to choose a decision that optimizes a single performance criterion. Although deliberately opting for less than the best possible leaves one open to charges of capriciousness, indecision, or foolhardiness, the incessant optimizer may be criticized as being restless, insatiable, or intemperate.[3] Just as moderation may tend to stabilize and temper cognitive behavior, deliberately backing away from strict optimality may provide protection against antisocial consequences. Moderation in the short run may turn out to be instrumentally optimal in the long run.

Even in the light of these considerations, substantive rationality retains a strong appeal, especially because it provides a systematic solution methodology, at least for single decision makers. One of the practical benefits of optimization is that by choosing beforehand to adopt the option that maximizes expected utility, the decision maker has completed the actual decision making – all that is left is to solve or search for that option (for this reason, much of what is commonly called decision theory may more accurately be characterized as search theory). This fact can be exploited to implement efficient search procedures, especially with concave and differentiable utility functions, and is a computational benefit of such enormous value that one might be tempted to adopt substantive rationality primarily because it offers a systematic and reliable means of finding a solution.

1.2.2 Procedural rationality

If we were to abandon substantive rationality, what justifiable notion of reasonableness could replace it? If we were to eschew optimization and its attendant computational mechanisms, how would solutions be systematically identified and computed? These are significant questions, and there is no single good answer to them. There is, however, a notion of rationality that has evolved more or less in parallel with the notion of substantive rationality and that is relevant to psychology and computer science.

Definition 1.4
Decision makers who make choices by following specific rules or procedures are said to be **procedurally rational** (Simon, 1986). □

[3] As Epicurus put it: "Nothing is enough for the man to whom enough is too little."

For an operational definition of procedural rationality, we turn to Simon:

> The judgment that certain behavior is "rational" or "reasonable" can be reached only by viewing the behavior in the context of a set of premises or "givens." These givens include the situation in which the behavior takes place, the goals it is aimed at realizing, and the computational means available for determining how the goals can be attained. (Simon, 1986, p. 26)

Under this notion, a decision maker should concentrate attention on the quality of the *processes* by which choices are made, rather than directly on the quality of the outcome. Whereas, under substantive rationality, attention is focused on *why* decision makers should do things, under procedural rationality attention is focused on *how* decision makers should do things. Substantive rationality tells us where to go, but not how to get there; procedural rationality tells us how to get there, but not where to go. Substantive rationality is viewed in terms of the outcomes it produces; procedural rationality is viewed in terms of the methods it employs.

Procedures are often heuristic. They may involve *ad hoc* notions of desirability, and they may simply be rules of thumb for selective searching. They may incorporate the same principles and information that could be used to form a substantively rational decision, but rather than dictating a specific option, the criteria are used to guide the decision maker by identifying patterns that are consistent with its context, goals, and computational capabilities.[4] A fascinating description of heuristics and their practical application is found in Gigerenzer and Todd (1999). Heuristics are potentially very powerful and can be applied to more complex and less well structured problems than traditional utility maximization approaches. An example of a procedurally rational decision-making approach is a so-called *expert system*, which is typically composed of a number of rules that specify behavior in various local situations. Such systems are at least initially defined by human experts or authorities.

The price for working with heuristics is that solutions cannot in any way be construed as optimal – they are functional at best. In contrast to substantively rational solutions, which enjoy an absolute guarantee of maximum success (assuming that the model is adequate – we should not forget that "experts" defined these models as well), procedurally rational solutions enjoy no such guarantee.

A major difference between substantive rationality and procedural rationality is the capacity for self-criticism, that is, the capacity for the decision maker to evaluate its own performance in terms of coherence and consistency. Self-criticism will be built into substantive rationality if the criteria used to establish optimality can also be used

[4] A well-known engineering example of the distinction between substantive rationality and procedural rationality is found in estimation theory. The so-called Wiener filter (Wiener, 1949) is the substantively rational solution that minimizes the mean-square estimation error of a time-invariant linear estimator. However, the performance of the Wiener filter is often approximated by a heuristic, called the LMS (least-mean-square) filter and developed by Widrow (1971). Whereas the Wiener filter is computed independently of the actual observations, the Widrow filter is generated by the observations. The Wiener filter requires that all stochastic processes be stationary and modeled to the second order; the Widrow filter relaxes those constraints. Both solutions are extremely useful in their appropriate settings, but they differ fundamentally.

to define the search procedure.[5] By contrast, procedural rationality does not appear to possess a self-policing capacity. The quality of the solution depends on the abilities of the expert who defined the heuristic, and there may be no independent way to ascribe a performance metric to the solution from the point of view of the heuristic. Of course, it is possible to apply performance criteria to the solution once it has been identified, but such *post factum* criteria do not influence the choice, except possibly in conjunction with a learning mechanism that could modify the heuristics for future application. While it may be too strong to assert categorically that heuristics are incapable of self-criticism, their ability to do so on a single trial is at least an open question.

Substantive rationality and procedural rationality represent two extremes. On the one hand, substantive rationality requires the decision maker to possess a complete understanding of the environment, including knowledge of the total preference orderings of itself and all other agents in the group. Any uncertainty regarding preferences must be expressed in terms of expectations according to known probability distributions. Furthermore, even given complete understanding, the decision maker must have at its disposal sufficient computational power to identify an optimal solution. Substantive rationality is highly structured, rigid, and demanding. On the other hand, procedural rationality involves the use of heuristics whose origins are not always clear and defensible, and it is difficult to predict with assurance how acceptable the outcome will be. Procedural rationality is amorphous, plastic, and somewhat arbitrary.

1.2.3 Bounded rationality

Many researchers have wrestled with the problem of what to do when it is not possible or expedient to obtain a substantively rational solution due to informational or computational limitations. Simon identified this predicament when he introduced the notion of **satisficing**.[6]

> Because real-world optimization, with or without computers, is impossible, the real economic actor is in fact a satisficer, a person who accepts "good enough" alternatives, not because less is preferred to more, but because there is no choice. (Simon, 1996, p. 28)

To determine whether an alternative is "good enough," there must be some way to evaluate its quality. Simon's approach is to determine quality according to the criteria used for substantive rationality, and to evaluate quality against a standard (the aspiration level) that is chosen more or less arbitrarily. Essentially, one continues searching for an optimal choice until an option is identified that meets the decision maker's aspiration level, at which point the search may terminate.

[5] This will be the case if the optimality existence proof is constructive. A non-constructive example, however, is found in information theory. Shannon capacity is an upper bound on the rate of reliable information transmission, but the proof that an optimal code exists does not provide a coding scheme to achieve capacity.

[6] This term is actually of ancient origin (*circa* 1600) and is a Scottish variant of satisfy.

The term "satisficing," as used by Simon, comprises a blend of the two extremes of substantive and procedural rationality and is a species of what he termed **bounded rationality**. This concept involves the exigencies of practical decision making and takes into consideration the informational and computational constraints that exist in real-world situations.

There are many excellent treatments of bounded rationality (see, e.g., Simon (1982a, 1982b, 1997) and Rubinstein (1998)). Appendix A provides a brief survey of the mainstream of bounded rationality research. This research represents an important advance in the theory of decision making; its importance is likely to increase as the scope of decision-making grows. However, the research has a common theme, namely, that if a decision maker could optimize, it surely should do so. Only the real-world constraints on its capabilities prevent it from achieving the optimum. By necessity, it is forced to compromise, but the notion of optimality remains intact. Bounded rationality is thus an approximation to substantive rationality, and remains as faithful as possible to the fundamental premises of that view.

I also employ the term "satisficing" to mean "good enough." The difference between the way Simon employs the term and the way I use it, however, is that satisficing *à la* Simon is an approximation to being best (and is constrained from achieving this ideal by practical limitations), whereas satisficing as I use it is treats being good enough as the ideal (rather than an approximation).

This book is not about bounded rationality. Rather, I concentrate on evaluating the appropriateness of substantive and procedural rationality paradigms as models for multi-agent decision making, and provide an alternative notion of rationality. The concepts of boundedness may be applied to this alternative notion in ways similar to how they are currently applied to substantive rationality, but I do not develop those issues here.

1.3 Middle ground

Substantive rationality is the formalization of the common sense idea that one should do the best thing possible and results in perhaps the strongest possible notion of what should constitute a reasonable decision – the only admissible option is the one that is superior to all alternatives. Procedural rationality is the formalization of the common sense idea that, if something has worked in the past, it will likely work in the future and results in perhaps the weakest possible notion of what should constitute a reasonable decision – an option is admissible if it is the result of following a procedure that is considered to be reliable. Bounded rationality is a blend of these two extreme views of rational decision making that modifies the premises of substantive rationality because of a lack of sufficient information to justify strict adherence to them.

Instead of merely blending the two extreme views of rational decision making, however, it may be useful to consider a concept of rationality that is not derived from either