Satisficing Games and **Decision Making**

With applications to engineering and computer science

Wynn C. Stirling

Brigham Young University

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typefaces Times 10.5/14 pt. and Helvetica Neue *System* LaTeX 2$_\varepsilon$ [TB]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data
Stirling, Wynn C.
Satisficing games and decision making : with applications to engineering and computer science / Wynn C. Stirling.
p. cm.
Includes bibliographical references and index.
ISBN 0 521 81724 2
1. Decision making. 2. Artificial intelligence. 3. Human-computer interaction. I. Title.
T57.95 .S73 2003
003′.56 – dc21 2002073606

ISBN 0 521 81724 2 hardback

Contents

Figures

Tables

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

the doctrine of rational choice or heuristic procedures. Kreps seems to express a desire along these lines when he observes that:

> . . . the real accomplishment will come in finding an interesting middle ground between hyperrational behaviour and too much dependence on *ad hoc* notions of similarity and strategic expectations. When and if such a middle ground is found, then we may have useful theories for dealing with situations in which the rules are somewhat ambiguous. (Kreps, 1990, p. 184)

Is there really a middle ground, or is the lacuna between strict optimality and pure heuristics bridgeable only by forming an *ad hoc* hybrid of these extremes? If a non-illusory middle ground does exist, it is evident that few have staked formal claims to any of it. The literature involving substantive rationality (bounded or unbounded), particularly in the disciplines of decision theory, game theory, optimal control theory, and operations research, is overwhelmingly vast, reflecting many decades of serious research and development. Likewise, procedural rationality, in the form of heuristics, rule-based decision systems, and various *ad hoc* techniques, is well-represented in the computer science, social science, and engineering literatures. Also, the literature on bounded rationality as a modification or blend of these two extremes is growing at a rapid pace. Work involving rationality paradigms that depart from these classical views, however, is not in evidence.

One of the goals of this book is to search not only for middle ground but for new turf upon which to build. In doing so, let us first examine a "road map" that may guide us to fruitful terrain. The map consists of desirable attributes of the notion of rationality we seek.

A-1 *Adequacy*: satisficing, or being "good enough," is the fundamental desideratum of rational decision makers. We cannot rationally choose an option, even when we do not know of anything better, unless we know that it is good enough. Insisting on the best and nothing but the best, however, can be an unachievable luxury.

A-2 *Sociality*: rationality must be defined for groups as well as for individuals in a consistent and coherent way, such that both group and individual preferences are accommodated. Group rationality should not be defined in terms of individual rationality nor vice versa.

These attributes represent a general relaxing of substantive rationality. Liberation from maximization may open the door to accommodating group as well as individual interests, while still maintaining the integrity supplied by adherence to principles. The attributes also bring rigor to procedural rationality, since they move away from purely *ad hoc* methods and insist on the capacity for self-criticism.

1.3.1 Adequacy

Adequacy is a harder concept to deal with than optimality. Achieving the summit of a mountain is a simple concept that does not depend upon the valley below. By contrast,

getting high enough to see across the valley depends upon the valley as well as the mountain. Optimality can be considered objective and is abstracted from context, but adequacy is subjective, that is, it is context dependent. Abstractification is powerful. It transforms a messy real-world situation into a clean mathematical expression that permits the power of calculus and probability theory to be focused on finding a solution. The advantages of abstractification are enormous and not lightly eschewed, and their appeal has fundamentally changed the way decision-making is performed in many contexts. But Zadeh, the father of fuzzy logic, suggests that always insisting on optimality is shooting beyond the mark, and that a softer notion of what is reasonable must be considered.

> Not too long ago we were content with designing systems which merely met given specifications . . . Today, we tend, perhaps, to make a fetish of optimality. If a system is not the "best" in one sense or another, we do not feel satisfied. Indeed, we are apt to place too much confidence in a system that is, in effect, optimal by definition . . .
>
> At present, no completely satisfactory rule for selecting decision functions is available, and it is not very likely that one will be found in the foreseeable future. Perhaps all that we can reasonably expect is a rule which, in a somewhat equivocal manner, would delimit a set of "good" designs for a system. (Zadeh, 1958)

A clear operational definition for what it means to be satisficing, or good enough, must be a central component of the notion of rationality that we are seeking. Zadeh reminds us that no such notion is likely to be a panacea, and any definition we offer is subject to criticism and must be used with discretion. Indeed, decision making is inherently equivocal, as uncertainty can never be completely eliminated.

To make progress in our search for what it means to be good enough, we must be willing to relax the demand for strict optimality. We should not, however, abandon the criteria that are used to define optimality, but only the demand to focus attention exclusively on the optimal solution. We certainly should not contradict the notion of optimality by preferring options that are poor according to the optimality criteria over those that comply with the criteria. The goal is to give place to a softer notion of rationality that accommodates, in a formal way, the notion of being good enough.

To maintain the criteria of optimality but yet not insist on optimality may seem paradoxical. If we know what is best, what possible reason could there be for not choosing it? At least a partial answer is that optimization is an ideal that serves to guide our search for an acceptable choice, but not necessarily to dictate what the final choice is. For example, when I drive to work my criterion is to get there in a timely manner, but I do not need to take the quickest route to satisfy the criterion. Strict optimality does not let me consider any but the very best route.

It is not irrational, in the view of some philosophers, for people not to optimize. Slote, for example, argues that it is reasonable not only to settle for something that is less than the best, but that such a situation may actually be preferred by a rational

decision maker. That is, one may willfully and rationally eschew taking the action that maximizes utility.

> Defenders of satisficing claim that it sometimes makes sense not to pursue one's own greatest good or desire-fulfillment, but I think it can also be shown that it sometimes makes sense deliberately to reject what is *better* for oneself in favor of what is *good and sufficient* for one's purposes. Those who choose in this way demonstrate a modesty of desire, a kind of moderation, that seems intuitively understandable, and it is important to gain a better understanding of such moderation if we wish to become clear, or clearer, about common-sense, intuitive rationality. (Slote, 1989, pp. 1–2; emphasis in original)

The gist of Slote's argument is that common sense rationality differs from optimizing views of rationality in a way analogous to the difference between common sense morality and utilitarian views of deontology. According to this latter view, what one is morally permitted to do, one is morally required to do. Similarly, substantive rationality requires one to optimize if one is able to do so. Slote argues that, just as utilitarian deontology prohibits decision makers from acting supererogatorily, that is, of doing more than is required or expected, optimizing views of rationality prohibit one from achieving less than one is capable of achieving. But common sense morality permits supererogation, and common sense rationality permits moderation.

Although Slote criticizes optimization as a model for behavior, he does not provide an explicit criterion for characterizing acceptable other-than-optimal activity. While an explicit criterion may not be necessary in the human context, when designing artificial agents, the designer must provide them with some operational mechanism to govern their decision-making if they are to function in a coherent way. Perhaps the weakest notion of rationality that would permit such activity is an operational notion of being "good enough."

One way to establish what it means to be good enough is to specify minimum requirements and accept any option that meets them. This is the approach taken by Simon. He advocates the construction of "aspiration levels" and to halt searching when they are met (Simon, 1955). Although aspiration levels at least superficially establish minimum requirements, this approach relies primarily upon experience-derived expectations. If the aspiration is too low, something better may needlessly be sacrificed, and if it is too high, there may be no solution. It is difficult to establish an adequate practically attainable aspiration level without first exploring the limits of what is possible, that is, without first identifying optimal solutions – the very activity that satisficing is intended to circumvent.[7] Furthermore, such an approach is susceptible to the charge that defining "good enough" in terms of minimum requirements begs the question, because the only way seemingly to define minimum requirements is that they are good enough.

[7] The decision maker may, however, be able to adjust his or her aspirations according to experience (see Cyert and March (1992)), in which case it may be possible to adopt aspiration levels that are near-optimal. Even so, however, there may be no way to determine how far one is away from the optimal solution without searching directly for it.

For single-agent low-dimensional problems, specifying the aspirations may be non-controversial. But, with multi-agent systems, interdependence between decision makers can be complex, and aspiration levels can be conditional (what is satisfactory for me may depend upon what is satisfactory for you).

Satisficing via aspiration levels involves making a tradeoff between the cost of continuing to search for a better solution than one currently has and the adequacy of the solution already in hand. That is, for any option under consideration, the decision maker makes a choice between accepting the option and stopping the search or rejecting the option and continuing the search. Making decisions in this way is actually quite similar to the way decisions are made under substantive rationality; it is only the stopping rule that is different. Both approaches rank-order the options and stop when one is found with acceptably high rank. With optimality, the ranking is relative to other options, and searching stops when the highest-ranking option is found. With aspiration levels, the ranking is done with respect to an externally supplied standard, and searching stops when an option is found whose ranking exceeds this threshold.

What aspiration levels and optimization have in common is that the comparison operation is *extrinsic*, that is, the ranking of a given option is made with respect to attributes that are not necessarily part of the option. In the case of optimization, comparisons are made relative to other options. In the case of aspiration levels, comparisons are made relative to an externally supplied standard. Under both paradigms, an option is selected or rejected on the basis of how it compares to things external to itself. Also, both rank-order comparisons and fixed-standard comparisons are global, in that each option is categorized in the option space relative to all other options.

Total ordering, however, is not the only way to make comparisons, nor is it the most fundamental way. A more primitive approach is to form dichotomies, that is, to define two distinct (and perhaps conflicting) sets of attributes for each option and either to select or reject the option on the basis of comparing these attributes. Such dichotomous comparisons are *intrinsic*, since they do not necessarily reference anything not directly relating to the option.

Whereas extrinsic decisions are of the form: either select Hamburger A or select Hamburger B (presumably on the basis of appearance and cost), intrinsic decisions are of the form: either select Hamburger A or reject Hamburger A, with a similar decision required for Hamburger B. The difference is that, under the extrinsic model, one would combine appearance and cost into a single utility that could be rank-ordered, but under the intrinsic model, one forms the binary evaluation of appearance versus cost. If only one of the hamburgers passes muster, the problem is resolved. If you conclude that neither hamburger's appearance is worthy of the cost, you are justified in rejecting them both. If you think both are worthy but you must choose only one, then you either may appeal to a more sophisticated (e.g., extrinsic) decision paradigm, or you may include additional criteria and try again, or you may make a random choice between the options. Suppose that Hamburger A costs more than Hamburger B, but is also much

larger and has more trimmings. By the intrinsic criteria, if you view both as being worth the price, then whatever your final choice, you at least get a good hamburger – you get your money's worth.

Dichotomies are the fundamental building blocks of everyday personal choices. Attached to virtually every nontrivial option are attributes that are desirable and attributes that are not desirable. To increase quality, one usually expects to pay more. To win a larger reward, one expects to take a greater risk. People are naturally wont to evaluate the upside versus the downside, the pros versus the cons, the pluses versus the minuses, the benefits versus the costs. One simply evaluates tradeoffs option by option – putting the gains and the losses on the balance to see which way it tips. The result of evaluating dichotomies in this way is that the benefits must be at least as great as the costs. In this sense, such evaluations provide a distinct notion of being good enough.

Definition 1.5
An option is **intrinsically rational** if the expected gains achieved by choosing it equal or exceed the expected losses, provided the gains and losses can be expressed in commensurable units. □

Definition 1.6
An option is **intrinsically satisficing** if it is intrinsically rational. □

By separating the positive (gain) and negative (loss) attributes of an option, I explicitly raise the issue of commensurability. It should be noted, however, that traditional utility theory also involves the issue of commensurability at least implicitly, since utility functions typically involve both benefits and costs, which are often summed or otherwise combined together to form a single utility function (for example, when forming a utility function for automobiles, positive attributes might be performance and reliability and negative attributes might be purchase and operating costs). Often such attributes can be expressed in, say, monetary units, but this is not always the case. Nevertheless, decision makers are usually able to formulate some rational notion of commensurability by appropriating or inventing a system of units. The issue was put succinctly by Hardin: "Comparing one good with another is, we usually say, impossible because goods are incommensurable. Incommensurables cannot be compared. Theoretically, this may be true; but in real life incommensurables *are* commensurable. Only a criterion of judgment and a system of weighing are needed" (Hardin, 1968, emphasis in original). Since my formulation of rationality requires explicit comparisons of attributes, the choice of units becomes a central issue and will be discussed in detail in subsequent chapters.

Intrinsic rationality is a weaker notion than substantive rationality, but it is more structured than procedural rationality. Whereas substantive rationality may be characterized

as an attitude of "nothing but the best will do" and procedural rationality may be characterized as an attitude of "it has always worked before," intrinsic rationality may be characterized as an attitude of "getting what you pay for." Substantive rationality assures optimality but is rigid. Procedural rationality is efficient but amorphous. Intrinsic rationality is ameliorative and flexible. There can be only one substantively rational option (or an equivalence class of them) for a given optimality criterion, and there can be only one procedurally rational option for a given procedure,[8] but there can be several intrinsically rational options for a given satisficing criterion.

The quality of a substantively rational option will be superior to all alternatives, according to the criteria used to define it. The quality of a procedurally rational option may be difficult to assess, since no explicit criteria are required to define it. The quality of intrinsically rational options may be uneven, since options that provide little benefit but also little cost may be deemed satisficing. Thus, intrinsic satisficing can be quite different from satisficing *à la* Simon.

My justification for using the term "satisficing" is that it is consistent with the issue that motivated Simon's original usage of the term – to identify options that are good enough by directly comparing attributes of the options to a standard. This usage differs only in the standard used for comparison. Whereas Simon's standard is extrinsic (attributes are compared to an externally supplied aspiration level), my standard is intrinsic (the positive and negative attributes of each option are compared to each other). If minimum requirements are readily available, however, it is certainly possible to define satisficing in a way that conforms to Simon's original idea.

Definition 1.7
An option is **extrinsically satisficing** if it meets minimum standards that are already supplied. □

Combining intrinsic and extrinsic satisficing is one way to remove some of the unevenness of intrinsic satisficing.

Definition 1.8
An option is **securely satisficing** if it is both intrinsically and extrinsically satisficing. □

It will not be assumed that minimum standards can always be specified. But if they are, it will be assumed that they employ a rationale that is compatible with that used to define gains and losses. If minimum standards are not available, the decision maker must still attempt to evaluate the unevenness of intrinsically satisficing solutions.

[8] With heuristics such as satisficing *à la* Simon, however, there may be multiple options that satisfy an extrinsic satisficing criterion, and the agent need not terminate its search after finding only one of them.

This issue will be discussed in detail in Chapter 5. Throughout the remainder of this book, the term satisficing will refer solely to intrinsic satisficing unless stated otherwise. It will be assumed that gains and losses can be defined, and that these attributes can be expressed in units that permit comparisons.

1.3.2 Sociality

Competition, which is the instinct of selfishness, is another word for dissipation of energy, while combination is the secret of efficient production. (Edward Bellamy, *Looking Backward* (1888))

Self-interested human behavior is often considered to be an appropriate metaphor in the design of protocols for artificial decision-making systems. With such protocols, it is often taken for granted that each member of a community of decision makers will try

. . . to maximize its own good without concern for the global good. Such self-interest naturally prevails in negotiations among independent businesses or individuals . . . Therefore, the protocols must be designed using a *noncooperative, strategic* perspective: the main question is what social outcomes follow given a protocol which *guarantees that each agent's desired local strategy is best for that agent – and thus the agent will use it.* (Sandholm, 1999, pp. 201, 202; emphasis in original)

When artificial decision makers are designed to function in a non-adversative environment, it is not obvious that it is either natural or necessary to restrict attention to noncooperative protocols. Decision makers who are exclusively focused on their own self-interest will be driven to compete with any other decision maker whose interests might possibly compromise their own. Certainly, conflict cannot be avoided in general, but conflict can just as easily lead to collaboration as to competition. Rather than head-to-head competition, Axelrod suggests that a superior approach is to look inward, rather than outward, and evaluate one's performance relative to one's own capabilities, rather than with respect to the performance of others.

Asking how well you are doing compared to how well the other player is doing is not a good standard unless your goal is to destroy the other player. In most situations, such a goal is impossible to achieve, or is likely to lead to such costly conflict as to be very dangerous to pursue. When you are not trying to destroy the other player, comparing your score with the other's score simply risks the development of self-destructive envy. A better standard of comparison is how well you are doing relative to how well someone else could be doing in your shoes. (Axelrod, 1984, p. 111)

This thesis is born out by the Axelrod Tournament (Axelrod, 1984), in which a number of game theorists were invited to participate in an iterated Prisoner's Dilemma[9]

[9] The Prisoner's Dilemma, to be discussed in detail in Section 8.1.3, involves two players who may either cooperate or defect. If one player cooperates and the other defects, the one who defects receives the best payoff while the one who cooperates receives the worst payoff. If both defect, they both receive the next-to-worst payoff, and if both cooperate, they both receive the next-to-best payoff (which is assumed to be better than the next-to-worst payoff).

tournament. The winning strategy was Rapoport's *tit-for-tat* rule: start by cooperating, then play what the other player played the previous round. What is interesting about this rule is that it always loses in head-to-head competition, yet wins the overall best average score in round-robin play. It succeeds by eliciting cooperation from the other players, rather than trying to defeat them.

Cooperation often involves *altruism*, or the notion that the benefit of others is one's ultimate goal. This notion is in contrast to *egoism*, which is the doctrine that the ultimate goal of every individual is to benefit only himself or herself. The issue of egoism versus altruism as an explanation for human behavior has captured the interest of many researchers (Sober and Wilson, 1998; Mansbridge, 1990a; Kohn, 1992). As expressed by Sober and Wilson:

> Why does psychological egoism have such a grip on our self-conception? Does our everyday experience provide conclusive evidence that it is true? Has the science of psychology demonstrated that egoism is correct? Has Philosophy? All of these questions must be answered in the negative . . . The influence that psychological egoism exerts far outreaches the evidence that has been mustered on its behalf . . . Psychological egoism is hard to disprove, but it also is hard to prove. Even if a purely selfish explanation can be imagined for every act of helping, this doesn't mean that egoism is correct. After all, human behavior also is consistent with the contrary hypothesis – that some of our ultimate goals are altruistic. Psychologists have been working on this problem for decades and philosophers for centuries. The result, we believe, is an impasse – the problem of psychological egoism and altruism remains unsolved. (Sober and Wilson, 1998, pp. 2, 3)

Peirce, also, is skeptical of egoism as a viable explanation for human behavior:

> Take, for example, the doctrine that man only acts selfishly – that is, from the consideration that acting in one way will afford him more pleasure than acting in another. This rests on no fact in the world, but it has had a wide acceptance as being the only reasonable theory. (Peirce, 1877)

It is not my intent to detail the arguments regarding egoism versus altruism as explanations for human behavior; such an endeavor is best left to psychologists and philosophers. But, if the issue is indeed an open question, then it would be prudent to refrain from relying exclusively on a rationality model based solely on self-interest when designing artificial entities that are to work harmoniously, and perhaps altruistically, with each other and with humans.

One of the possible justifications for adopting self-interest as a dominant paradigm for artificial decision-making systems is that it is a simple and convenient principle upon which to build a mathematically based theory. It allows the decision problem to be abstracted from its context and expressed in unambiguous mathematical language. With this language, utilities can be defined and calculus can be employed to facilitate the search for the optimal choice. The quintessential manifestation of this approach to decision making is von Neumann–Morgenstern game theory (von Neumann and Morgenstern, 1944). (See Appendix B for a brief summary of game theory basics.)

Under their view, game theory is built on one basic principle: individual self-interest – each player must maximize its own expected utility under the constraint that other players do likewise. For two-person zero-sum games (see Definition B.6 in Appendix B), individual self-interest is perhaps the only reasonable, non-vacuous principle – what one player wins, the other loses. Game theory insists, however, that this same principle applies to the general case. Thus, even in situations where there is the opportunity for group as well as individual interest, only individually rational actions are viable: if a joint (that is, for the group) solution is not individually rational for some decision maker, that self-interested decision maker would not be a party to such a joint action. This is a rigid stance for a decision maker to take, but game theory brooks no compromises that violate individual rationality.

Since many decision problems involve cooperative behavior, decision theorists are tempted to define notions of group preference as well as individual preference. The notion of group preference admits multiple interpretations. Shubik describes two, neither of which is entirely satisfactory to game theorists (in subsequent chapters I offer a third): "Group preferences may be regarded either as derived from individual preferences by some process of aggregation or as a direct attribute of the group itself" (Shubik, 1982, p. 109). Of course, not all group scenarios will admit a harmonious notion of group preference. It is hard to imagine a harmonious concept of group preference for zero-sum games, for example. But, when there are joint outcomes that are desirable for the group to obtain, the notion of group interest cannot be ignored.

One way to aggregate a group preference from individual preferences is to define a "social-welfare" function that provides a total ordering of the group's options. The fundamental issue is whether or not, given arbitrary preference orderings for each individual in a group, there always exists a way of combining these individual preference orderings to generate a consistent preference ordering for the group. In an landmark result, Arrow (1951) showed that no social-welfare function exists that satisfies a set of reasonable and desirable properties, each of which is consistent with the notion of self-interested rationality and the retention of individual autonomy (this theorem, known as Arrow's impossibility theorem, is discussed in more detail in Section 7.3).

The Pareto principle provides a concept of social welfare as a direct attribute of the group.

Definition 1.9

A joint (group) option is a **Pareto equilibrium** if no single decision maker, by changing its decision, can increase its level of satisfaction without lowering the satisfaction level of at least one other decision maker. □

As Raiffa has noted, however, the Pareto equilibrium can be equivocal.

It seems reasonable, does it not, that the group *should* choose a Pareto-optimal act? Otherwise there would be alternative acts that at least some would prefer and no one would "disprefer". Not too long

ago this principle seemed to me unassailable, the one solid cornerstone in an otherwise swampy area. I am not so sure now, and I find myself in that uncomfortable position in which the more I think the more confused I become.

One can argue that the group by its very existence should have a common bond of interest. If the members disagree on fundamentals (here, on probabilities and on utilities) they ought to thrash these out independently, arrive at a compromise probability distribution and a compromise utility function, and use these in the usual Bayesian manner. (Raiffa, 1968, p. 233, emphasis in original)

Adopting this latter view would require the group to behave as a *superplayer*, or, as Raiffa puts it, the "organization incarnate," who functions as a higher-level decision maker. Shubik refers to the practice of ascribing preferences to a group as a subtle "anthropomorphic trap" of making a shaky analogy between individual and group psychology. He argues that, "It may be meaningful . . . to say that a group 'chooses' or 'decides' something. It is rather less likely to be meaningful to say that the group 'wants' or 'prefers' something" (Shubik, 1982, p. 124). Shubik criticizes the view of the group as a superplayer capable of ascribing preferences according to some sort of group-level welfare function as being too narrow in scope to "contend with the pressures of individual and factional self-interest." Although Raiffa also rejects the notion of a superplayer, he still feels "a bit uncomfortable . . . somehow the group entity is more than the totality of its members" (Raiffa, 1968, p. 237).

Arrow expresses a similar discomfort: "All the writers from Bergson on agree on avoiding the notion of a social good not defined in terms of the values of individuals. But where Bergson seeks to locate social values in welfare judgments by individuals, I prefer to locate them in the actions taken by society through its rules for making social decisions" (Arrow, 1951, p. 106). Although Arrow does not tell us how such rules should be defined or, once defined, how they should be implemented, his statement nevertheless expresses the notion that societies may possess structure that is more complicated than can be expressed via individual values.

Perhaps the source of this discomfort is that, while individual rationality may be appropriate for environments of perfect competition, it loses much of its power in more general sociological settings. As Arrow noted, the use of the individual rationality paradigm is "ritualistic, not essential" (Arrow, 1986). What is essential, however, is that any useful model of society accommodate the various relationships that exist between the agents. But achieving this goal should not require artifices such as the aggregation of individual interests or the creation of a superplayer.[10] While such approaches may be recommended by some as ways to account for group interests, they may also manifest the limits of the substantive rationality paradigm.

Nevertheless, game theory, which relies exclusively upon self-interest, has been a great success story for economics and has served to validate the assumption of

[10] Margolis (1990) advocates a "dual-utilities" approach, comprising a social utility and a private utility, with the decision maker allocating resources to achieve a balance between the two utilities. Margolis' approach eschews the substantive rationality premise, and is very much in the same spirit as the approach I develop in subsequent chapters.

substantive rationality in many applications. This success, however, does not imply that self-interest is the only principle that will lead to credible models of economic behavior, it does not imply the impossibility of accommodating both group and individual interests in some meaningful way, and it does not imply that individual rationality is an appropriate principle upon which to base a theory of artificial decision-making entities.

Game theory provides a systematic way of analyzing behavior where the consequences of one player's actions depend on the actions taken by other players. Even single-agent decision problems can be viewed profitably as games against nature, for example. The most common solution concepts of game theory are dominance and Nash equilibria.

Definition 1.10
A joint option is a **dominant equilibrium** if each individual option is best for the corresponding player, no matter what options the other players choose. □

Definition 1.11
A joint option is a **Nash equilibrium** if, were any single decision maker to change its decision, it would reduce its level of satisfaction. □

A dominant equilibrium corresponds to the ideal situation of all players being able simultaneously to maximize their own satisfaction. This is a rare situation, even for games where coordination is possible. Nash equilibrium is a much more useful concept, but not all games possess pure (that is, non-random) Nash equilibria. Nash (1950) established, however, that if random play is permitted where each player makes decisions according to a probability rule (a mixed strategy), then at least one Nash equilibrium can be found for a finite-player, finite-action game.

In contrast to Pareto equilibria, Nash equilibria is a strictly selfish concept, hence is not amenable to cooperative play. But an individually rational player would have no incentive to agree to a Pareto equilibrium if that solution did not assure at least as much satisfaction as the player could be guaranteed of receiving were it to ignore completely the interests of the other players.

Definition 1.12
The minimum guaranteed benefit that a player can be assured of achieving is its **security level.** □

Furthermore, a subgroup of players would have no incentive to agree to a joint solution unless the total benefit to the subgroup were at least as great as the minimum that could be guaranteed to the subgroup – its security level – if it acted as a unit (assuming transferable utilities which may be be reapportioned via side payments).

The **core** of an N-person game is the set of all solutions that are Pareto equilibria and at the same time provide each individual and each possible subgroup with at least their security levels (the concept of the core is discussed in more detail in Section 7.1). Unfortunately, the core is empty for many interesting and nontrivial games.

An empty core exposes the ultimate ramifications of a decision methodology based strictly on the maximization of individual expectations. There may be no way to meet all of the requirements that are imposed by strict adherence to the dictates of individual rationality. There are many ways to justify solutions that are not in the core, such as accounting for bargaining power based on what a decision maker calculates that it contributes to a coalition by joining it (e.g., the Shapley value), or by forming coalitions on the basis of no player having a justified objection against any other member of the coalition (e.g., the bargaining set).

I do not criticize the rationale for these refinements to the theory, nor do I criticize the various extra-game-theoretical considerations that may govern the formation of coalitions, such as friendship, habits, fairness, etc. I simply point out that to achieve a reasonable solution it may be necessary to go beyond the strict notion of maximizing individual expectations and employ ancillary assumptions that temper the attitudes and abilities of the decision makers. There are many such ingenious and insightful solution concepts but, as Shubik notes,

> Each solution probes some particular aspect of rational individuals in mutual interaction. But all of them have had to make serious compromises. Inevitably, it seems, sharp predictions or prescriptions can only be had at the expense of severely specialized assumptions about the customs or institutions of the society being modeled. The many intuitively desirable properties that a solution ought to have, taken together, prove to be logically incompatible. (Shubik, 1982, p. 2)

This observation cuts to the heart of the situation: under von Neumann–Morgenstern game theory, any considerations of customs and peculiarities of the collective that are not explicitly modeled by the individual utility functions are extra-game-theoretic and must be accommodated by some sort of add-on logic. Much of the ingenuity and insight associated with game theory may lie in devising ways to force these considerations into the framework of individual rationality. While this practice may be appropriate for the *analysis* of human behavior, it is less appropriate for the *synthesis* of artificial decision-making entities, since any such idiosyncratic attributes must be an explicit part of the decision logic, not merely a *post factum* explanation for anomalous behavior. I suggest, however, that the problem is more fundamental than simply accounting for idiosyncrasies.

The critical issue, in my view, has to do with the structure of the utility functions. Before articulating this point, let me first briefly summarize utility theory as it is employed in mathematical games. Utility theory was developed as a mathematical way to encode individual preference orderings. It is built on a set of axioms that describe how a “rational man” would express his preference between two alternatives in a consistent

Table 1.1: Payoff array for a two-player game with two strategies each

	X_2	
X_1	s_{21}	s_{22}
s_{11}	$(\pi_1(s_{11}, s_{21}), \pi_2(s_{11}, s_{21}))$	$(\pi_1(s_{11}, s_{22}), \pi_2(s_{11}, s_{22}))$
s_{12}	$(\pi_1(s_{12}, s_{21}), \pi_2(s_{12}, s_{21}))$	$(\pi_1(s_{12}, s_{22}), \pi_2(s_{12}, s_{22}))$

way.[11] An expected utility function is a mathematical expression that is consistent with the preferences and conforms to the axioms. Since, in a game-theoretic context, an individual's preferences are generally dependent upon the payoffs (expected utilities) that obtain as a result of the individual's strategies and of the strategies available to others, an individual's expected utility function must be a function not only of the individual's own strategies, but of the strategies of all other individuals. For example, consider a game involving two players, denoted X_1 and X_2, such that each player has a strategy set consisting of two elements, that is, X_1's set of strategies is $S_1 = \{s_{11,}s_{12}\}$ and X_2's set of strategies is $S_2 = \{s_{21}, s_{22}\}$ (for this single-play game, strategies are synonymous with options). X_1's expected utility function would be a function $\pi_1(s_{1j}, s_{2k})$, $j, k = 1, 2$. Similarly, X_2's expected utility function is of the form $\pi_2(s_{1j}, s_{2k})$. Thus, each individual computes its expected utility as a function of both its own strategies and the strategies of the other players. These expected utilities may then be juxtaposed into a payoff array, and solution concepts may be devised to define equilibrium strategies, that is, strategies that are acceptable for all players. Table 1.1 illustrates the payoff array for a two-player game with two strategies each.

The important thing to note about this structure is that *it is not until the expected utilities are juxtaposed into an array so that the expected utility values for all players can be compared that the actual "game" aspects of the situation emerges.* It is the juxtaposition that reveals possibilities for conflict or coordination. These possibilities are not explicitly reflected in the individual expected utility functions by themselves. In other words, although the individual's expected utility is a function of other players' strategies, *it is not a function of other players' preferences.* This structure is completely consistent with exclusive self-interest, where all a player cares about is its personal benefit as a function of its own and other players' strategies, without any regard for the benefit to the others. Under this paradigm, the only way the preferences of others factor into an individual's decision-making deliberations is to constrain behavior to limit the amount of damage they can do to oneself. Pareto equilibria notwithstanding, a true notion of group rationality is not a logical consequence of individual rationality.

[11] This is not to say that the axioms cannot be generalized to deal with group preferences, but the theory has not been developed that way.

Table 1.2: Payoff matrix in ordinal form for the Battle of the Sexes game

	S	
H	*D*	*B*
D	(4, 3)	(2, 2)
B	(1, 1)	(3, 4)

Key: 4 = best; 3 = next best; 2 = next worst; 1 = worst

Luce and Raiffa summarize the situation succinctly:

> . . . general game theory seems to be in part a sociological theory which does not include any sociological assumptions . . . it may be too much to ask that any sociology be derived from the single assumption of individual rationality. (Luce and Raiffa, 1957, p. 196)

Often, the most articulate advocates of a theory are also its most insightful critics. Yet, such criticism is not often voiced, even by advocates of game theory as a model of human behavior. For example, consider the well-known Prisoner's Dilemma game (see Section 8.1.3). This game is of interest because possibilities for both cooperation and conflict are present, yet under the paradigm of individual rationality, only the joint conflict solution (the Nash equilibrium) is rational.

The Prisoner's Dilemma game may be an appropriate model of behavior when (a) the opportunity for exploitation exists, (b) cooperation, though possible, incurs great risk, and (c) defection, even though it offers diminished rewards, protects the participant from catastrophe. Many social situations, however, possess a strong cooperative flavor with very little incentive for exploitation. One prototypical game that captures this feature is the Battle of the Sexes game (Bacharach, 1976) to be discussed in detail in Section 8.1.2. This is a game involving a man and a woman who plan to meet in town for a social function. She (S) prefers to go to the ballet (B), while he (H) prefers the dog races (D). Each also prefers to be with the other, however, regardless of venue. The classical way to formulate this game is via a payoff matrix, as given in Table 1.2 in ordinal form, with the payoff pairs representing the benefits to H and S, respectively.

Rather than competing, these players wish to cooperate, but they must make their decisions without benefit of communication. Both players lose if they make different choices, but the choices are not all of equal value to the players. This game has two Nash equilibria, (D, D) and (B, B).

One of the perplexing aspects of this game is that it does not pay to be altruistic (deferring to the venue preferred by the other), since, if both participants did, they would each receive the worst outcome. Nor does it pay for both to be selfish (demanding the venue preferred by oneself) – that guarantees the next worst outcome for each player. The best and next-best outcomes obtain if one player is selfish and the other altruistic.

It seems that a way to account for the preferences of others when specifying one's own preferences would be helpful, but there is no obvious way to do this within the conventional structure.

Taylor (1987) addresses the issue of accounting for the interests of others by introducing a formal notion of altruism that involves transforming the game to a new game according to a utility array whose entries account for the payoffs to others as well as to oneself. Taylor suggests that the utility functions be expressed as a weighted average of the payoffs to oneself and to others. By adjusting the weights, a player is able to take into consideration the payoffs of others.

Taylor's form of altruism does not distinguish between the state of *actually relinquishing* one's own self-interest and the state of *being willing to relinquish* one's own self-interest under the appropriate circumstances. To relinquish unconditionally one's own self-interest is a condition of *categorical* altruism – a decision maker unconditionally modifies its preferences to accommodate the preferences of others. A purely altruistic player would completely replace its preferences with the preferences of others. A state of being willing to modify one's preferences to accommodate others if the need arises is a state of *situational* altruism. Here, a decision maker is willing to accommodate, at least to some degree, the preferences of others in lieu of its own preferences if doing so would actually benefit the other, but otherwise retains its own preferences intact and avoids needless sacrifice.

Categorical altruism may be too much to expect from a decision maker who has its own goals to pursue. However, the same decision maker may be willing to engage, at least to a limited degree, in a form of situational altruism. Whereas it is one thing for an individual to modify its behavior if it is sure that doing so will benefit another individual (situational), it is quite another thing for an individual to modify its behavior regardless of its effect on the other (categorical). In the Battle of the Sexes, If H knew that S had a very strong aversion to D (even though S would be willing to put up with those extremely unpleasant surroundings simply to be with H and thus receive her second-best payoff), H might then prefer B to D. But if S did not have a strong aversion to D then H would stick to his preference for D over B (in Section 8.1.2 I introduce situational altruism into this game).

This example seems to illustrate Arrow's claim that, when the assumption of perfect competition fails, "the very concept of [individual] rationality becomes threatened, because perceptions of others and, in particular, of their rationality become part of one's own rationality" (Arrow, 1986). Arrow has put his finger on a critical weakness of individual rationality: it does not provide a way to incorporate another's rationality into one's own rationality without seriously compromising one's own rationality. I do not assert that, under the theoretical framework of conventional game theory, it is impossible to formulate theoretical models of social behavior that go beyond individual interests and accommodate situationally altruistic tendencies while at the same time preserving individual preferences. However, the extant literature does not provide such

a theory. I assert that it will be difficult to develop such a theory that remains compatible with the principle of individual rationality.

There are many ways to introduce categorical altruism into the design of artificial decision makers. One approach is to modify the decision maker's utility function to become a function of the group's payoff. In effect, the player is "brainwashed" into substituting group interests for its personal interests. Then, when acting according to its supposed self-interest, it is actually accommodating the group (Wolpert and Tumer, 2001). A somewhat similar, though less radical, approach is taken by Glass and Grosz (2000) and Cooper et al. (1996), who attempt to instill a social consciousness into agents, rewarding them for good social behavior by adjusting their utility functions with "brownie points" and "warm glow" utilities for doing the "right thing."

It is certainly possible for human altruists to interpret their sacrifice as, ultimately, a benefit to themselves for having made another's good their own (motivated, possibly, by such "pure" altruistic attributes as duty and love, or perhaps by "impure" altruistic attributes such as the sense of power that derives from having helped another (Mansbridge, 1990b)), but it seems less appropriate to ascribe such anthropomorphic interpretations (or motives) to artificial decision-making entities. While, granting that it is possible for a decision maker to suppress its own preferences in deference to others by redefining its own expected utility to be maximized, doing so is little more than a device for co-opting individual rationality into a form that can be interpreted as unselfish. Such a device only simulates attributes of cooperation, unselfishness, and altruism while maintaining a regime that is competitive, exploitive, and avaricious. Altruism, springing from whatever motive in man or machine, may often be accommodated in multi-agent relationships, but it does not follow that it can be accommodated within a regime that recognizes self-interest as the primary basis for rational decision making.

Social choice theory is another multi-agent formalism that has been widely studied. Like game theory, this theory has been developed largely on the foundation of individual rationality. For example, Harsanyi defines a social welfare function as a positive linear combination of individual utilities where each individual utility in this combination is a mapping of group options to individual utility. Each player then proceeds according to the substantively rational paradigm by maximizing its expected utility subject to any constraints that are relevant (Harsanyi, 1977).

The social welfare function modifies the decision maker's stance from a consideration of purely selfish preferences to a consideration of what are termed *moral* (or social) preferences, and gives weight to the interests of each participant. However, the sequence of mappings from group options to individual utilities and then from individual utilities to a group utility provides a very constrained linkage between one decision maker's preferences (for itself or for the group) and another decision maker's utilities and may not deal adequately with the rich diversity of interconnections that can exist in

multi-agent groups. Furthermore, such mappings constitute unconditional (categorical) changes to the individual's utilities.

One of the characteristics of perhaps all societies, except for those that are either completely anarchic or completely dictatorial, is that group and individual preferences are woven together in a complex fabric that is virtually impossible to decompose into constituent pieces that function independently. Exclusive self-interest simply does not capture the richness and complexity of functional societies. On the other hand, to relinquish fundamental control over individual preferences and focus primarily on the preferences of the group as a whole may not be feasible, since individuals can be asked to make unreasonable sacrifices that place them in extremely disadvantageous situations. This suggests that functional societies must achieve some sort of equilibrium that is flexible enough to accommodate the preferences of both the individual and the group. Such an approach would be consistent with Levi's dictum that

> . . . principles of coherent or consistent choice, belief, desire, etc. will have to be weak enough to accommodate a wide spectrum of potential changes in point of view. We may not be able to avoid some fixed principles, but they should be as weak as we can make them while still accommodating the demand for a systematic account. (Levi, 1997, p. 24)

Achieving, or at least approximating, equilibria involving both group and individual preferences is an essential condition for a system of autonomous artificial decision makers if they are to be representative of human groups. Obtaining such a state, however, requires a generalized notion of utility that seamlessly combines group and individual interests, even though it is individuals, and individuals only, who make the decisions. Such a utility theory must therefore be based on a notion of preference that allows group preferences to influence individual preferences and thereby to influence individual actions.

Accommodating group preferences must not leave the individual open to an unintentional or unacceptable degree of self-sacrifice. Thus, there must be a clear means of evaluation so that the individual can control the amount of compromise it is willing to consider. In other words, the individual must possess a means for self-control.

Heuristics offer no such capability. Under procedural rationality, once an individual adopts a rule that accommodates any form of compromise that exposes it to self-sacrifice, it becomes difficult to control the extent of its commitment without knowing beforehand the strategies of the other participants.

If one is willing to consider an option that is not strictly in its own best interest, one must be able to add some friction to the slippery slope of compromise. One way to do this is to adopt a satisficing stance, where satisficing is applied to the group as well as to the individual. Whereas optimization is strictly an individual concept, satisficing can be a social, as well as an individual, concept. For any group of decision makers, if the group and each of its members is willing to compromise sufficiently, there will

exist a joint option that is good enough for the group as a whole and good enough for each member of the group according to their individual standards (this claim is made explicit in Section 7.2). This does not mean, of course, that the decision makers are obligated to accept this compromise option. It means only that it exists.

The remainder of this book explores the concept of intrinsic rationality, instantiated at both the individual and group levels, as a means of achieving an equilibrium of shared preferences and acceptable compromises. Intrinsic satisficing requires the specification of two general types of preferences – gains and losses. For a single-agent decision, it is conceptually straightforward to place each of the relevant attributes into one of these categories. When dealing with more than one decision maker, however, the interactions between them are not so readily categorized. Relationships are interconnected and conditional: one decision maker's gains and losses may affect other decision maker's gains and losses. Furthermore, the interconnections that exist between players must be at the level of preference interconnections, rather than action interconnections, as they are usually expressed in conventional game theory. The method of characterizing these preferences must be exhaustive, so that all possible relationships between decision makers can be represented, but at the same time it must be parsimonious, so that it is not more complex than it needs to be.

The central message of this book is that exclusive self-interest, coupled with strict optimality, is indeed an "excess of reasonableness." Self-interest is not the bedrock of rationality. Decision making, especially in group settings, can be ameliorated by relaxing the demands for optimization in its various forms (global maximization, constrained maximization, minimax, and even such "boundedly rational" approaches such as Simon's aspiration-level satisficing).