1

Sociality

One can argue very persuasively that the weakest link in any chain of argument should not come at the beginning.

— *Howard Raiffa*
Decision Analysis (Addison-Wesley, 1968)

Decision making is perhaps the most fundamental intellectual enterprise. Indeed, the word *intelligent* comes from the Latin roots *inter* (between) + *legĕre* (to choose). The study of motives and methods regarding how decisions might and should be made has long captured the interest of philosophers and social scientists and, more recently, of engineers and computer scientists. An important objective of such studies is to establish a framework within which to define rational behavior and solution concepts that result in appropriate choices. The development of formal theories of decision making, however, has proven to be a challenging and complex task. The reason is simple: Every nontrivial decision problem involves multiple stakeholders. A stakeholder is any entity that has an interest in the consequences of a decision, whether or not it has direct control over the decision. Unless the interests of all stakeholders coincide perfectly (a rare occurrence), conflicts will exist. The central challenge of any theory of decision making, therefore, is how to make choices in the presence of conflicting stakeholder interests.

The way a group of stakeholders deals with conflict is a function of its sociality: Conflict can result in either competition or cooperation. At one extreme (speaking anthropomorphically), each member of a group views others as adversaries and treats them antagonistically. At the other extreme, it views others as partners and treats them synergistically. On the one hand, others are viewed as opponents that can only constrain achieving one's selfish desires; on the other hand, others are viewed as teammates with whom to pursue common objectives that cannot be achieved effectively alone. Members of a group typically

fall between these two extremes, displaying mixed motives, as they balance opportunities to benefit themselves regardless of the expense to others with opportunities to benefit others at possibly their own expense.

Any systematic account of rational decision making requires that a mathematical model be specified that defines the stakeholders, their possible alternatives, their preferences over the alternatives, and their concepts of rational behavior.[1] Game theory, as developed by von Neumann and Morgenstern (1944) is perhaps the best known such mathematical model. In this treatment, we restrict attention to finite strategic (normal form) single-stage noncooperative games.[2] Formally, a game consists of two or more autonomous stakeholders, or players, each of whom has a finite set of pure strategies (deterministic actions in a single-stage context) from which it may choose one action to instantiate. An action profile is an array of actions, one for each player. These profiles constitute the outcomes, or consequences, of the game. Each player also possesses a preference ordering over the outcomes. Typically, these preference orderings are expressed in terms of numerical valuations, called payoffs or utilities, that define the benefits (either ordinally or cardinally) to the players.

Representing the decision problem as a mathematical game permits the players to strip the decision problem from its context and to examine it dispassionately from the point of view of possible actions and outcomes. Most developers of such models have faithfully adhered to Occam's razor and have resisted the introduction of complicating factors that are not deemed essential. When defining a game, however, it is imperative also to consider what some have termed Einstein's razor: "It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience" (Einstein, 1934, p. 165).[3] Seemingly, everyone can agree with the first part of this dictum to keep things simple. But it is the latter injunction, not to surrender adequate representation, that is perhaps more difficult to accommodate.

Once the stakeholders and actions are specified, two fundamental elements remain to be defined when formulating a multistakeholder decision problem: (a) the structure of the preference orderings over the outcomes and (b) the notions of rationality that are used to formulate solution concepts.

[1] Henri Poincaré observed that "mathematics is a language by which no indistinct, obscure, or indefinite things can be expressed" (cited in Mérö (1998, p. 230).)
[2] For many scenarios, a multistage game can be described by a series of single-stage games, particularly when the key issue of concern is coordination.
[3] This quote has sometimes been reworded into such variants as "Everything should be made as simple as possible, but not simpler."

By far the most prevalent assumption employed by decision theory when considering preference orderings is also the most simple: A preference ordering over outcomes is well defined for each individual stakeholder. Arrow (1951) put it succinctly: "It is assumed that each individual in the community has a definite ordering of all conceivable social states, in terms of their desirability to him. . . . It is simply assumed that the individual orders all social states by whatever standards he deems relevant" (p. 17). According to this view, each stakeholder's preference ordering is completely and immutably defined by the payoffs before the stakeholders engage in the act of making choices. Such a preference ordering is *categorical*: It unconditionally defines the stakeholder's valuation system under all circumstances, regardless (at least ostensibly) of the valuations of other stakeholders.

The second fundamental element of a decision model is the concept of rational behavior that governs the way stakeholders use the information at their disposal. The simplest possible rationality model, which is also the most widely used, is that each member of a group will restrict interest to its own benefit and will act in a way that achieves its best possible outcome. This is the doctrine of *individual rationality*. As observed by Tversky and Kahenman (1986),

> The assumption of [individual] rationality has a favored position in economics. It is accorded all of the methodological privileges of a self-evident truth, a reasonable idealization, a tautology, and a null hypothesis. Each of these interpretations either puts the hypothesis of rational action beyond question or places the burden of proof squarely on any alternative analysis of belief and choice. The advantage of the rational model is compounded because no other theory of judgment and decision can ever match it in scope, power, and simplicity. (p. 89)

The mathematical structure of categorical preference orderings and the logical structure of individual rationality are ideally matched to each other. Harsanyi (1977) articulated this point as follows: "Because all values and objectives in which the players are interested have been incorporated into their payoff functions, our formal analysis of any given game must be based on the assumption that each player has only one interest in the game – to maximize his own payoff" (p. 13). Given categorical preference orderings, the only compatible notion of rationality is self-interest, since the preferences are restricted to, and only to, individual welfare. Conversely, given individual rationality, any structure other than categorical preferences would extend interest beyond the self. With these structures, it is possible to formulate precise definitions of rational behavior for the members of a group.

What is lacking with this structure, however, is a concept of preference for the group and, hence, a concept of rational behavior for the group. One reason for the lack of focus on group preferences is that since the group, as an entity,

does not possess the capability to enforce a decision even if it had a preference, the relevance of a group preference ordering is problematic. Furthermore, as Arrow's (1951) impossibility theorem establishes, it is not generally possible to define a preference ordering for a group by simply aggregating individual preferences of its members without violating a set of arguably reasonable and desirable properties. Thus, although it is the joint decision of the members of the group that defines the outcome and, hence, the benefits to the individuals within the group, the group itself does not possess a preference ordering. As argued by Luce and Raiffa (1957), "the notion of group rationality is neither a postulate of the model nor does it appear to follow as a logical consequence of individual rationality" (p. 193). Consequently, classical decision theory has proceeded by making assumptions about individual preferences only and then using those preferences to deduce information about the choices (but not the values) of a group.

The classical game-theoretic approach of focusing exclusively on individual preferences may be justified when the members of a group fit the model; that is, when they really are able to define their preferences categorically and are motivated by, and only by, self-interest. But that is an extreme situation; it is an abstraction that must be justified in application, not merely taken as a tenet of a classical doctrine to be applied uncritically. Arrow (1986) clearly delimits the context in which this model applies:

> Rationality in application is not merely a property of the individual. Its useful and powerful implications derive from the conjunction of individual rationality and other basic concepts of neoclassical theory – equilibrium, competition, and completeness of markets.…When these assumptions fail, the very concept of rationality becomes threatened, because perceptions of others and, in particular, their rationality become part of one's own rationality. (p. 203)

Despite Arrow's (1986) concern, the classical model of categorical preferences and individual rationality is routinely applied in contexts where, in addition to competition, opportunities for cooperation, compromise, and unselfishness are present. The application of the classical model in such contexts can lead to paradoxes and dilemmas that, although interesting and even charming, may be evidence that the model does not adequately account for the social relationships that exist among the members of the group.

Maslow observed that if the only tool you have is a hammer, you tend to see every problem as a nail. The classical game theory model may be a good hammer to drive the nail of competitive and market-driven decision making, but it may not be the best tool to model scenarios where more sophisticated social relationships exist. Although this tool has been effective in economic contexts, as the applications of multistakeholder decision making expand into

other domains, its limitations become more pronounced. In his musing about the history of decision making, Shubik (2001) commented on the limitations of the classical approach to multistakeholder decision making:

> Economic man, operations research man and the game theory player were all gross simplifications. They were invented for conceptual simplicity and computational convenience in models loaded with implicit or explicit assumptions of symmetry, continuity, and fungibility in order to allow us (especially in a pre-computer world) to utilize the methods of calculus and analysis. Reality was placed on a bed of Procrustes to enable us to utilize the mathematical techniques available. (p. 4)

There are essentially two ways to address the limitations of the classical model: One may retrofit the old bed of categorical preferences and individual rationality, or one may create a new framework that is designed to deal explicitly with social relationships that are more complex than competition, including cooperation, compromise, and negotiation. Pursuing the former approach simply means continuing to force a group with a complex social structure into the classical preference structure/rationality bed; with the latter approach, the goal is to make the bed a better fit for its occupant.

The central theme of this book is to present a new model that (a) permits individuals to modulate their preference orderings to accommodate the interests of others as well as themselves and (b) employs notions of rationality that simultaneously apply to both individuals and the group. This new model sits at the intersection of two diverse disciplines: the social sciences, including economics, psychology, sociology, and political science, and the engineering and computer science disciplines (including distributed artificial intelligence, intelligent control theory, and multiagent systems theory). Although they have much in common mathematically, these two disciplines have different application scenarios. In the social sciences, models are used for analysis – that is, to explain, predict, justify, and recommend choices for human societies – but for engineering and computer science, models are used for synthesis – that is, to design and construct artificially intelligent decision-making societies such as multiagent systems and distributed control systems that must function autonomously. Whereas models used in the social sciences to characterize human behavior are noncausal and serve only as approximations to reality, models used in the engineering and computer science contexts are causal – they dictate behavior and create reality. It is critical, therefore, that such models must be capable of explicitly accounting for complex social relationships when they exist (the analysis context) or when they are desired (the synthesis context).

The intent of this presentation is to supplement, rather than supplant, existing theory and methodology. However, since this theory challenges some of the closely held assumptions that have served for decades as the foundational tenets

of multistakeholder decision theory, it is important to review these assumptions, discuss their limitations, and lay the framework for going beyond them.

1.1 Classical theory

The doctrine of individual rationality has had enormous influence in the formulation of decision theories. One of the main justifications for this doctrine is its apparent consistency with the evolutionary theory of natural selection. The basic idea is that selfish characteristics evolve because they make the individual more fit for survival. Thus, natural selection promotes egoism. Furthermore, the argument goes, natural selection inhibits altruism, since helping others to survive would likely diminish the individual's own chances for surviving.

Nevertheless, the stubborn fact of ostensibly altruistic behavior in human society is readily observed. Examples abound of people sacrificing their own interests to benefit others. But sacrificing one's own interest to benefit another can also be viewed as fundamentally egoistical (people act in that way to feel good about themselves). At the end of the day, however, arguments that self-interest is the primary motive for human behavior are inconclusive. Sober and Wilson (1998) sum it up this way:

> 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. (pp. 2–3)

If the problem is unresolved, it would be prudent to refrain from relying exclusively on a rationality model that is based solely on self-interest, either when analyzing human behavior or, especially, when designing artificially intelligent entities that are intended to work harmoniously, and perhaps altruistically, with each other and with humans.

As a motive for action, self-interest is perhaps the most common justification that is invoked by decision theory for behavior in multistakeholder contexts. Without doubt, it is the simplest motive imaginable, since it takes into consideration only the benefit to the individual. Nevertheless, the very concept of self-interest, as a well-defined motive, has come under criticism. The key argument is that the concept is so simple that it is essentially vacuous. As argued by Holmes (1990), "The decision to group together sharply dissimilar motives under the single category of 'calculating self-interest' is said to involve an undesirable *loss of information* about rudimentary psychological and behavioral processes.

This is the essence of Macaulay's mocking remark that to discover self-interest behind an action is to say, with tautological banality, that 'a man had rather do what he had rather do' (Macaulay, 1978) [emphasis in original][4] (p. 269).

On the other hand, advocates of the self-interest concept argue that, while perhaps oversimplified, it nevertheless enables the construction of mathematical models to characterize behavior, and serves a prescriptive, even if not necessarily descriptive, role in the formation of a quantitatively based theory of value and behavior. As noted by Hogarth and Reder (1986),

> The role of [individual] rationality is to provide a principle (or "rationale") to mediate the relations between changes in one or more resource constraints and changes in the quantities of the relevant phenomena. This takes the form of the maintained hypothesis that *each of the individual decision makers behaves as if he or she were solving a constrained maximization problem* [emphasis added]. (p. 3)

The issue is further clouded by the existence of two flavors of self-interest: *narrow self-interest* and *enlightened self-interest*. With the former, the individual defines its preferences in accordance with its own welfare, and only its own welfare, regardless of the effect on others. With the latter, the individual defines its preferences in a way that improving the welfare of others ultimately improves its own welfare. This distinction may be important when considering the process of how one defines one's preference ordering, but once encoded into categorical preference orderings, self-interest is self-interest, regardless of the modifier, and the decision is made dispassionately according to whatever solution concept is applied.

There are no restrictions on the criteria that an individual uses to define its own self-interest. Although there is no explicit mechanism for an individual to consider the interests of others, the individual can always simulate the interests of others by substituting those interests in lieu of its own. As Sen (1990) observed:

> It is possible to define a person's interests in such a way that no matter what he does he can be seen to be furthering his own interests in every isolated act of choice. . . . No matter whether you are a single-minded egoist or a raving altruist or a class-conscious militant, you will appear to be maximizing your own utility in this enchanted world of definitions. (p. 19)

Once self-interest is redefined, a new game is created, and the issue devolves to considerations of which game is to be played. (In fact, one could randomly

[4] There is an interesting parallel between self-interest and survival of the fittest. "The semantic emptiness of the doctrine [survival of the fittest] was long ago exposed by asking the simple question, 'Fittest for what?' The only possible answer is 'fittest to survive,' which closes the circle and thereby reduces the statement to complete nonsense by making it read: 'Those survive who survive' " (Krutch, 1953, p. 85).

choose between the two games in a desperate attempt to reconcile one's conflicted interests, but such an approach would only obfuscate the issue further and delay a more comprehensive approach.) The only reasonable assumption from the point of view of classical game theory is that each player's payoffs defines its true preferences.

If an individual has concerns for the welfare of others as well as itself, restricting it to a categorical preference ordering severely limits its ability to characterize an enlarged sphere of interest. This restriction may be appropriate when self-interest is the operative attribute, but it does not easily account for such attributes as concern for others, willingness to cooperate, and moral principles. When observed behavior deviates from the behavior predicted by the model, it is necessary to invoke psychological and sociological attributes that, while not part of the mathematical model, are necessary to explain the deviations. They merely overlay the basic mathematical structure of a game and avoid or postpone a more profound solution, namely, the introduction of a model structure that explicitly accounts for complex social relationships and accommodates notions of rational behavior that extend beyond narrow self-interest.

1.2 Sociality models

Homans (1961) offers three criteria for behavior to qualify as *social*. First, an individual's actions must elicit some form of reward or punishment as a result of behavior by another individual. Second, behavior toward another individual must result in reward or punishment from that individual, not just a third party. Third, the behavior must be actual behavior, not just a norm of behavior.

1.2.1 Minimal sociality

Under the paradigm of individual rationality, the stakeholders are indifferent, at least ostensibly, to the welfare of others; they are manifestly egocentric. Even so, it can be the case that some notion of social behavior can emerge as a result of strategic reasoning, such as the attainment of a Nash equilibrium, a strategy profile such that if any player unilaterally deviates, then its payoff is reduced. A Nash equilibrium is a solution to a constrained optimization problem: Each player does the best for itself, assuming that all other players are similarly motivated. Such behavior, however, benefits only the individuals; benefit to the group is undefined. Thus, although the behavior is a minimal expression of sociality, a concept of group-level welfare is nonexistent.

Table 1.1. *The payoff matrix in ordinal form for the Prisoner's Dilemma game*

	X_2	
X_1	C	D
C	(3, 3)	(1, 4)
D	(4, 1)	(2, 2)

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

Example 1.1 Perhaps the most well known of all games is the Prisoner's Dilemma (PD). The players, denoted X_1 and X_2, each have two possible actions: cooperate (C) or defect (D). Conventionally, the game is defined by an ordinal payoff matrix of the form displayed in Table 1.1. This game serves as a model for situations where mutual cooperation is better than mutual noncooperation for both players, but unilaterally attempting to cooperate leaves one player vulnerable to exploitation by the other. If one player chooses to cooperate and the other defects, the one who attempts to cooperate receives the worst payoff, while the other receives the best payoff. The dilemma arises because mutual defection (D, D) is the unique Nash equilibrium, and, according to the doctrine of individual rationality, the players should adopt this pessimistic next-worst solution rather than the more optimistic next-best (Pareto optimal) mutual cooperation solution (C, C). □

Regardless of the choices that are made, there is no clearly defined notion of group preference for the Prisoner's Dilemma; that is, the preference of the group is viewed as a whole and not individually. Of course, an external party is free to ascribe a notion of group preference, such as arguing that the group as a whole is better off if both choose to cooperate, but such an exogenously ascribed notion would be arbitrary.

An argument can also be made that mutual cooperation can emerge as a result of repeatedly playing the PD game, and that such behavior can be viewed as a group-level notion of preference. If the game is played an indefinite or a random number of times, the incentive to defect can be overcome by the threat of punishment, and mutual cooperation can emerge as an equilibrium, as attested by the many experiments and contests that have been performed with this and other games (Axelrod, 1984). It is important to appreciate, however, that this

result is a consequence of learning and is not an intrinsic property of the model structure or of individual rationality. The players simply learn that, in the long run, it is in their better individual interest to cooperate. This result is relevant to the development of theories of learning and demonstrates how cooperation, and perhaps even notions of group preference, can evolve. Although such games serve as platforms with which to conduct important psychological experiments, the results do not alter the fact that the mathematical structure of categorical preferences and the logical structure of individual rationality do not accommodate an explicitly discernible notion of group preference and, hence, of group rationality.

One of the primary virtues of a mathematical model is that it provides a quantitative description of the values and preferences of the stakeholders. Ideally, such a model will strip the problem of all irrelevant and redundant issues and reduce it to its bare-bones mathematical essence. Rasmusen (1989) terms this no-fat modeling: "No-fat modeling is an extension of Occam's razor and the *ceteris paribus* assumption so fundamental to economics or, indeed, to any kind of analysis. The heart of the approach is to discover the simplest assumptions needed to generate an interesting conclusion: the starkest, barest, model that has the desired result" (pp. 14–15). This modeling assumption is compatible with the "hourglass" approach described by Slatkin (1980). According to Slatkin's approach, a complex problem is introduced, then reduced to a tractable mathematical model by stripping away all irrelevant issues, and finally, once a solution is obtained, expanded back into the original context for interpretation.

It is the last element of the hourglass approach: however, that is the most problematic. Many examples of social situations exist where classical game theory does an inconsistent job of explaining or predicting human behavior (e.g., the Prisoner's Dilemma, the Ultimatum game). Evidently, representing a social encounter by a set of categorical preference orderings can remove some meat along with the fat.

In an attempt to keep the meat on the bone, the field of behavioral economics has augmented the concept of self-interest to render it more psychologically realistic by incorporating notions of fairness and equity into the individuals' preference orderings. We illustrate this situation with the following example.

Example 1.2 The Ultimatum game has received great attention as a purported example of irrational behavior; that is, as a case where the players of the game are motivated by considerations other than maximizing their individual benefit. The setup of this two-player game is as follows: X_1, called the *proposer*, has access to a fortune, f, and offers X_2, called the *responder*, a portion $p \leq f$, and X_2 chooses whether or not to accept the offer. If X_2 accepts, then the two