Faculty Publications

2005-11-01

Social utility Functions-part I: theory

Wynn C. Stirling
wynn_stirling@byu.edu

Follow this and additional works at: <https://scholarsarchive.byu.edu/facpub>

Part of the [Electrical and Computer Engineering Commons](#)

Original Publication Citation

Stirling, W. C. "Social Utility Functions-Part I: Theory." *Systems, Man, and Cybernetics, Part C: Applications and Reviews*, IEEE Transactions on 35.4 (25): 522-32

BYU ScholarsArchive Citation

Stirling, Wynn C., "Social utility Functions-part I: theory" (2005). *Faculty Publications*. 342.
<https://scholarsarchive.byu.edu/facpub/342>

This Peer-Reviewed Article is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Faculty Publications by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.

Social Utility Functions—Part I: Theory

Wynn C. Stirling

Abstract—The dominant approaches to utility-based multiagent decision theory rely on the premise of individual rationality—the doctrine that each individual is committed to achieving the best outcome for itself, regardless of the effect doing so has on others. This fundamentally asocial concept is the basis of conventional von Neumann–Morgenstern (vN-M) utilities but is inadequate to characterize truly cooperative artificial systems. Social utility functions differ from conventional vN-M utilities in that they are functions of multiple decision-maker preferences, rather than actions, and thus permit individuals to expand their spheres of interest beyond the self. A logical basis for coherent reasoning in multiagent environments must obey exactly the same desiderata as do multivariate probability functions. By taking a dual utilities approach (one to account for effectiveness and one to account for efficiency), a new game-theoretic structure, called satisficing games, provides a decision-making procedure that accounts for both individual and group interest and presents a framework for the design of sophisticated multiagent societies.

Index Terms—Decision making, distributed control, game theory, intelligent systems, multiagent systems, probability theory, rationality, satisficing games.

I. INTRODUCTION

MULTIAGENT decision theory deals primarily with systems of decision makers, such as robots and other artificially intelligent agents, who must function without real-time supervision in their environments to accomplish the tasks for which they are designed. If such systems are to be understood, trusted, and confidently used by man, their operations must be *rational*; that is, they must conform to logical principles of organized and constructive behavior. Studies of rationality are often viewed as the purview of philosophy and the social sciences, and therefore are only marginally within the scope of engineering concern. This may be true when confining interest to single-agent behavior for, in that restricted domain, there is an obvious and seemingly incontrovertible notion of rational behavior that is often taken for granted: doing (or at least approximately doing) the best thing possible—optimizing. This is the doctrine of *individual rationality*. In a multiagent context, such a prescription for behavior is overly simplistic. Even in a cooperative environment, what is best for one agent may be injurious to others. Rarely will it be the case that a unique notion exists for what is best for the system as a whole as well as for each of its individual members. Individual rationality is the Occam's razor of social relationships: every agent is intent on doing the best thing for itself regardless of the effect doing so has on others.

The problem with individual rationality is that only individuals can optimize. If a group were to optimize its behavior, then it must act as if it were a superplayer or, as Raiffa puts it, the “organization incarnate” [1]. However, the resulting solution for the group would not necessarily be optimal, or even acceptable, for the individuals that compose the group. Optimization drives a wedge between individual and group interests. This is the conundrum of individual rationality: how to address conflicts between individuals and the group.

Philosophers and economists have long wrestled with the notion of rational behavior. Arrow noted that “Among the classical economists, such as Smith and Ricardo, rationality had the limited meaning of preferring more to less” [2]. Nozick describes a notion of *instrumental rationality* as “the effective and efficient pursuit of given goals” [3]. Harsanyi expands on this concept: “Rational behavior is simply behavior consistently pursuing some well defined goals, and pursuing them according to some well defined set of preferences or priorities” [4]. Modern decision theory has taken these general concepts to the extreme, treating rational behavior as synonymous with “optimal” behavior to the exclusion of virtually all other notions. As noted by Tversky and Kahneman [5, p. 89] “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.”

This same mentality applies to much of artificial agent design, where the imperative at least to approximate an “optimal” solution is often taken for granted. As noted by Weiss, “‘Intelligent’ indicates that the agents pursue their goals and execute their tasks such that they optimize some given performance measures” [6, p. 2]. Russell and Norvig are of this same mind: “In a sense, the MEU [maximum expected utility] principle could be seen as defining all of AI. All an intelligent agent has to do is calculate the various quantities, maximized utility over its actions, and away it goes” [7, p. 585]. Of course, as the complexity, size, time constraints, and exigencies of real-world multiagent decision making intensify, strict optimization may, by practical necessity, be replaced by approaches with a heuristic flavor. Concepts of bounded rationality [8]–[11], for example, take into consideration time and computational constraints, and seek to find an acceptable solution within the exigencies of the problem. Such methods are typically based on individual rationality and are not philosophically different from strict optimization.

Total reliance on the doctrine of individual rationality is questioned by psychologists; Sober and Wilson argue as follows: “Why does psychological egoism have such a grip on our self-conception? Does our everyday experience provide conclusive

Manuscript received June 4, 2004; revised August 27, 2004 and October 5, 2004. This paper was recommended by Associate Editor C. J. Malmberg.

The author is with the Electrical and Computing Engineering Department, Brigham Young University, Provo, UT 84602 USA (e-mail: wynn@ee.byu.edu).
Digital Object Identifier 10.1109/TSMCC.2004.843198

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” [12, pp. 2, 3].

Optimization is a strongly entrenched concept of the social sciences. It is a central doctrine of neoclassical economic theory, as developed by Bergson and Samuelson [13], [14], who formalized the notion of the “rational man” (*homo economus*). This doctrine has been embraced by the engineering community, also with great success. It is not the intent of this paper to join the argument that currently exists between the neoclassical economists and other schools of thought within that discipline (e.g., see [15]), except to say that, if the doctrine of individual rationality (and hence optimization) is subject to legitimate criticism as an analysis tool to predict and explain human social behavior, then it ought also to be held in question as a legitimate synthesis tool to be used by engineers as a foundational paradigm for designing artificial societies of multiagent systems who must function in ways that are understandable and acceptable to humans. Obviously, there are many applications where individual rationality provides an adequate foundation for multiagent decision making. However, since a model of rational behavior is the initial link in the chain of decision-making logic, we would do well to take heed of Raiffa's astute observation that “One can argue very persuasively that the weakest link in any chain of argument should not come at the beginning” [1, p. 130] and ensure that our commitment to individual rationality (and, hence, to optimization) is justified.

Almost four decades ago, Zadeh cautioned against unreflective or precipitant reliance upon optimality: “Today, we tend, perhaps, to make a fetish of optimality. If a system is not ‘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 . . . 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” [16]. While optimization may be an adequate concept with which to model egoism (despite Zadeh's caution), it may not be adequate for characterizing such sophisticated social concepts as cooperation, compromise, negotiation, and altruism. It may be fruitful to consider the following questions: Can decisions be rational if they are not based, ultimately, on some notion of optimality and, if so, what other metric or metrics of quality can be used to characterize rational decisions? If such questions can be affirmatively answered, then another question follows: Does such a concept of rationality provide a foundation for the design of multiagent systems that are capable of sophisticated social behavior?

This paper addresses these fundamental questions. Its goal is to present and justify an alternative notion of rational behavior

that removes the wedge between group and individual interests. To paraphrase Zadeh, it must provide a reasonable way to delimit the set of good, or more precisely, good enough, designs. As demanded by Tversky and Kahneman, it must shoulder the burden of proof to establish its reasonableness. To proceed, the most well-known theoretical approach to rational behavior in group settings—von Neumann–Morgenstern (vN-M) game theory—is briefly explored, and its limited ability to account for both group and individual interests is discussed. An alternative view of utility theory that focuses on the utility of preferences, rather than the utility of actions, and establishes a logical basis for coherent reasoning in multiagent contexts is presented. Using these social utility functions, a corresponding theory of multiagent decision making, termed satisficing games, is described, and it is shown how this alternative concept permits a resolution of the group/individual conundrum.

II. VON NEUMANN–MORGENSTERN GAME THEORY

A well-known theoretical tool for studying multiagent systems is vN-M game theory [17], which is the instantiation of individual rationality in multiagent settings. A common solution concept is for each player to compute the constrained optimal decision for itself under the assumption that all others are doing likewise. The result is a Nash equilibrium—a mutual condition such that, if any individual participant were to change its decision, its payoff would be reduced. This approach is attractive for at least three reasons.

- 1) It is simple in that it makes only the minimal sociological assumption that all players are egoistic and wish to maximize their own satisfaction.
- 2) It is self-enforcing, since players who are committed to maximizing their own benefit do not need an external party to dictate their behavior.
- 3) It is amenable to systematic design synthesis—optimization (unconstrained if possible and constrained if necessary).

An alternative candidate for rational behavior is a Pareto optimal solution—a mutual condition such that, if any individual were to change its decision in an attempt to improve its payoff, the payoff for some other individual would decrease. Unfortunately, Pareto optimality is not a generally accepted solution concept since it is not self-enforcing.

Consider the Prisoner's Dilemma game, undoubtedly the most well-known and most intensely studied of all games. This is a mixed-motive game that often serves as a model of social behavior when opportunities for both cooperation and exploitation exist. This game comprises two players, X and Y , who may either cooperate (C) or defect (D), with rewards defined by the payoff matrix given in Table I. There are two well-defined notions of behavior: the Nash equilibrium (D, D) (next worst for both) and the Pareto optimal solution (C, C). The Nash solution is the individually rational one, at least for one-off play; it protects against exploitation of the self while at the same time presents the opportunity to exploit another's vulnerability. On the other hand, the Pareto optimal decision, which clearly improves the individual payoffs over

the Nash solution, also leaves one vulnerable to exploitation, and hence must be eschewed by an individually rational player. However, the Pareto solution could be deemed rational if either 1) the players were to possess information that is not explicitly present in the payoff array (such as the results of previous play) or 2) the spheres of interest of the players were to extend beyond the self. (For in-depth discussions of this game, see, e.g., [18], and [19]).

One way to account for the extension of interest beyond the self is to form a notion of group rationality as a companion to individual rationality. Such a concept is troubling to game theorists, however, because, as put by Luce and Raiffa, “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” [20, p. 193]. Shubik identifies two general ways to account for group rationality: “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” [21, p. 108]. Both of these ways, however, present significant problems.

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. This is a “bottom-up” approach, whereby the interest of the group is a composite of the interests of the individuals. 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’s impossibility theorem [22] proves that no social-welfare function exists that satisfies a set of reasonable and desirable properties (such as transitivity), each of which is consistent with the notion of self-interested rationality and the retention of individual autonomy.

Pareto optimality provides a concept of group interest as a direct attribute of the group. However, this notion of rational behavior falls short of a viable solution concept for individually rational decision makers, since such a player would not consent to reducing its own satisfaction simply to benefit another—it is not self-enforcing. Adopting this view would require the group to behave as a superplayer who functions as a higher level decision maker who can compel the players to conform to a concept of group interest that may not be compatible with individual interests. This is an example of a “top-down” approach, whereby the behavior of the individuals are imposed by a central authority.

While vN–M game theory provides a powerful analysis tool to explain and predict behavior, the strength of this approach is also its weakness—individual rationality—since it admits only a rudimentary and limited notion of interagent influence. This limitation is well known to game theorists. Over four decades ago, Luce and Raiffa observed that “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” [20, p. 196]. It is often the case that the most articulate advocates of a theory are also its most insightful critics.

TABLE I
PAYOFF MATRIX IN ORDINAL FORM FOR THE PRISONER’S
DILEMMA GAME. KEY: 4 = BEST; 3 = NEXT BEST;
2 = NEXT WORST; AND 1 = WORST

	Y	
X	C	D
C	(3, 3)	(1, 4)
D	(4, 1)	(2, 2)

Neither top-down nor bottom-up structures are adequate to solve the group/individual interest conundrum.

III. UTILITIES

To overcome the limitations of individual rationality for multiagent decision making, we propose to return to the headwaters of preference formulation, rather than somewhere downstream, and reexamine the fundamental structure of the way preferences are expressed, and hence the way utilities are defined. A *preference pattern* is a total ordering of all possible actions available to the decision maker that reflects the desirability of the consequences of the actions that can be undertaken. Let $\{\succeq, \cong\}$ denote a binary ordering relationship between elements of a set U , meaning *{is at least as good as, is equivalent to}* in terms of the consequences that obtain to the decision maker. Since it is a total ordering, it is reflexive ($\forall u \in U : u \succeq u$), antisymmetric ($\forall u, u' \in U : u \succeq u' \text{ and } u' \succeq u \implies u \cong u'$), transitive: $\forall u, u', u'' \in U : u \succeq u', u' \succeq u'' \implies u \succeq u''$, and linear ($\forall u, u' \in U : u \succeq u' \text{ or } u' \succeq u$).

Preference orderings are ordinal in the sense that, while it may be the case that $u \succeq u'$, the ordering relation does not specify how much, or to what degree, u is superior to u' . If the degree of superiority is important to the design of the decision system, then it will be necessary to establish cardinal relationships that are consistent with the ordinal relationships. Such cardinal orderings are accomplished by means of a utility function. A fundamental utility existence theorem states that, given a preference pattern $\{\succeq, \cong\}$ for a set of options U , and a set of reasonable assumptions (i.e., that no option is infinitely desirable or infinitely undesirable and that convex combinations of preferences preserve the ordering—see [23] and [24]), then there exists a real-valued function f such that, for $u, u' \in U$

$$u \succeq u' \text{ if and only if } f(u) \geq f(u'). \quad (1)$$

The function f is called a *utility*. Furthermore, if f is a utility, then any positive affine transformation of the form $f'(u) = \alpha f(u) + \beta$ for $\alpha > 0$ is also a utility. A proof of this theorem and methods for the construction of utilities are discussed in numerous places (see, e.g., [17] and [23]–[26]).

The standard way to define utilities is for each individual member of a group to order the options available to it as a function of the actions that others, as well as itself, may take. Such utilities are called vN–M utilities. Consider a system comprising N agents, each with its option set $U_i, i = 1, 2, \dots, N$. Let $\mathbf{u} = (u_1, u_2, \dots, u_N)$ denote a vector of options such that $u_i \in U_i$. Then, player i ’s utility function is a real-valued function $f_i(\mathbf{u})$.

For example, Table I comprises the utility functions for the Prisoner's Dilemma (e.g., $f_x(D, C) = 4$ and $f_y(D, C) = 1$).

Although vN–M utilities correspond to preference orderings for each player conditioned on the possible actions that other players might instantiate, they do not take into account the other player's attitudes about instantiating those actions. Consider the Prisoner's Dilemma. The fact that (D, C) yields the worst consequence for Y does not influence X 's evaluation of that action vector as the best for itself. This structure is entirely consistent with individual rationality: when defining its preferences, each player considers its interest in a social "vacuum," in the sense that it does not let the preferences of others influence its preference ordering. It is not until the utilities of all players are juxtaposed into a payoff array that opportunities for conflict and cooperation become evident. Essentially, vN–M utilities are asocial—they possess no capacity for extending one's sphere of interest beyond the self. This fundamental structure of vN–M utility theory makes it difficult to use to create a decision-making methodology that is able to accommodate nonegoistic behavior. Attempting to accommodate altruism within the doctrine of individual rationality is an extremely difficult enterprise; a standard approach is for the players to redefine their individual payoffs to make them reflect not only their own interests, but the interests of others as well [27]–[29]. With these modified payoffs, the players then seek to maximize their reward by optimizing in the standard way. Although it is possible to suppress one's preferences in deference to others by redefining one's own payoff, 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 fundamentally competitive, exploitive, and avaricious. Sen exposed the weakness of such attempts by noting that "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" [30, p. 19].

The question before us is this: Can we create a rational multiagent solution methodology that accommodates sophisticated social relationships, such as cooperation, compromise, negotiation, and altruism? In contrast to the asocial utilities that model only individual rationality, we desire to identify social utilities that can model complex relationships. To address this question, we must define an alternative notion of utility that applies to both groups and individuals. To guide our search, it is helpful to recognize the parallels between epistemology and praxeology. Epistemology is the study of propositions on the basis of knowledge and belief regarding their content, and praxeology is "the study of such actions as are necessary in order to give practical effect to a theory or technique; the science of human conduct; the science of efficient action" [31]. Although the term praxeology does not currently enjoy widespread technical usage, it captures the intent of many designs, particularly those that involve the interaction of multiple entities. Whereas epistemology deals with the classification of propositions in terms of truth and belief, praxeology deals with the classification of options

in terms of their efficacy and practicality. Essentially, epistemology addresses the question "what to believe," and praxeology addresses the question "how to act." The obvious close connection between these two modes of endeavor (indeed, one could consider the "act" of "believing") suggests that they may be discussed in the same mathematical language. To develop this connection, let us first review the fundamental requirements of what may be termed epistemic preferences, then investigate analogous requirements for praxeic preferences, and finally express the two concepts under the same mathematical umbrella.

A. Epistemic Preferences

Suppose we wish to consider a finite set of propositions such that exactly one of them is true, but we are uncertain as to which is the true one. A *plausibility preference pattern* is a total ordering of the propositions in terms of truth support. When considering multiple random propositions, we must take into account the fact that the random phenomena may be linked by logical mechanisms. In the interest of brevity, we restrict our attention to two finite sets of random propositions U_i $i = 1, 2$, each with their individual plausibility preference patterns $\{\succeq_i, \cong_i\}$, $i = 1, 2$, meaning *is at least as plausible than, is as plausible as*. We also assume the existence of a joint plausibility preference pattern $\{\succeq_{12}, \cong_{12}\}$ that orders the plausibility of joint propositions $(u_1, u_2) \in U_1 \times U_2$. By the utility existence theorem, there exist utility functions f_1, f_2 , and f_{12} that correspond to the individual and joint plausibility preference patterns, respectively. We require that the joint and individual preference patterns satisfy two general coherence properties. The first is that sure loss must be avoided, and the second is that joint plausibility must not contradict individual conditional plausibility.

To illustrate the property of avoiding sure loss, let r and d be weather (w) conditions of rain and dry, respectively, and let c and p denote sky (s) conditions of cloudy and partly cloudy, respectively. Let $\{\succeq_w, \cong_w\}$ and $\{\succeq_s, \cong_s\}$ denote plausibility preference patterns for the weather and sky conditions, respectively, and let $\{\succeq_{ws}, \cong_{ws}\}$ denote the plausibility preference pattern for the joint weather/sky conditions. To avoid sure loss, suppose, say, $(r, c) \succeq_{ws} (d, c)$ and $(r, p) \succeq_{ws} (d, p)$; then, $r \succeq_w d$ must hold. To see why this must be so, suppose one were to enter a lottery to guess whether or not it will rain to win a prize of \$1 for a correct guess. One concept of a fair entry fee would be $\$p$ (where $0 \leq p \leq 1$), the plausibility of $(r, c) \vee (r, p)$, while another concept would be $\$q$ (where $0 \leq q \leq 1$), the plausibility of d . By paying an entry fee of $p + q$, one would be assured of winning \$1, regardless of the outcome. Now suppose that the disjunction $(r, c) \vee (r, p)$ is at least as plausible as the disjunction $((d, c) \vee (d, p))$, but d is at least as plausible as r . According to these two plausibility orderings, a fair entry fee would require $p \geq 1/2$ and $q \geq 1/2$. Thus, the player would be required to pay $p + q \geq 1$ to return \$1 for certain, resulting in a sure loss of $p + q - 1$. (This type of incoherency is sometimes called a "Dutch book.") Avoiding sure loss requires $p + q \leq 1$.

In general, let U_i , $i = 1, 2, \dots, N$ be a collection of finite sets of random propositions, each with elements $u_i \in U_i$, $i = 1, 2, \dots, N$, and suppose there exists a preference pattern $\{\succeq_i, \cong_i\}$ for each set and that there exists a joint preference

pattern $\{\succeq_{12\dots N}, \cong_{12\dots N}\}$ for the entire collection taken simultaneously. Then, to avoid sure loss, we require that

$$(u_1, \dots, u_i \dots u_N) \succeq_{12\dots N} (u_1, \dots, u'_i \dots u_N) \forall u_j \in U_j, \\ j \neq i \implies u_i \succeq_i u'_i.$$

To illustrate the second property, that joint plausibility must not contradict individual conditional plausibility, we assume the existence of conditional epistemic preference patterns that provide orderings for the plausibility of unknown propositions given known propositions. We write $u_i|u_j \succeq_i u'_i|u_j$, where the symbol “|” is termed the conditioning symbol, meaning that u_i is at least as plausible as u'_i , given that u_j is true. Again using the above climate example to illustrate, consider the conditional propositions $r|c$ and $d|c$, where $r|c$ is the proposition that it is raining, given that it is cloudy, etc. If, say, $r|c \succeq_w d|c$, then $(r, c) \succeq_{ws} (d, c)$ must hold. If this condition did not hold, then the fact that one happened merely to observe the cloud conditions could somehow influence the plausibility of rain. In general, this concept of coherency requires that

$$u_i|(u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_N) \\ \succeq_i u'_i|(u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_N) \\ \implies (u_1, \dots, u_i, \dots, u_N) \\ \succeq_{12\dots N} (u_1, \dots, u'_i, \dots, u_N).$$

This requirement means, given that the subvector $(u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_N)$ is true, that the plausibility of the joint event $(u_1, \dots, u_N)$ can be ascertained without separately evaluating the plausibility of u_i —it is sufficient to know the conditional plausibility of $u_i|(u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_N)$.

The coherency conditions of avoiding sure loss and avoiding contradictions, in addition to the total ordering requirement, ensure that the corresponding utility functions are adequate characterizations of epistemic uncertainty. Section III-C establishes a mathematical structure that guarantees these conditions.

B. Praxeic Preferences

A utility function provides a means of numerically evaluating options in terms of their efficacy. If we restrict attention to the consideration of individual preference, the only requirement on the utility function is that it satisfy (1). This structure is also adequate in the multiple-agent context, so long as we restrict to narrow self-interest. However, to extend beyond individual rationality, we must define utilities that accommodate more general preference patterns. One way to make this extension is for the individuals to have the ability to define their preferences conditioned on the preferences for action of others, rather than as functions directly of the actions of others as is done with vN–M utilities.

Consider the Prisoner’s Dilemma, and let $\{\succeq_x, \cong_x\}$ and $\{\succeq_y, \cong_y\}$ denote the individual preference patterns for X and Y , respectively. Now suppose that X were able to specify its conditional preference for its choice given that Y prefers its choice. For example, $D|C \succeq_x C|C$ means that X prefers D to C given that Y prefers C to D . By conditioning on preferences, rather than directly on actions, X can express selfishness, altruism, cooperation, or indifference. X would

express selfishness if $D|C \succeq_x C|C$ (the classical result under the vN–M regime). X would express altruism (a self-sacrifice to benefit another) if $C|D \succeq_x D|D$. X would express cooperation if $C|C \succeq_x D|C$. X would express indifference if its preferences were not a function of Y ’s preferences—for example, if $D|D \succeq_x C|D$ and $D|C \succeq_x C|C$.

The ability to specify conditional preferences as functions of the preferences of others (as well as one’s own) enlarges the sphere of a player from narrow self-interest to the genuine consideration of the interests of others and opens the door for expressions of group interest. For groups other than complete anarchies, that is, groups with at least some principles that relate the individuals to each other, a notion of group preference can be defined. Even for purely competitive situations such as zero-sum games, a fundamental principle exists: mutual opposition, resulting in a group preference to oppose each other. With purely coordinative games, the group preference is to function harmoniously. With mixed-motive games, the notion of group preference is highly context dependent.

Although we do not require the group praxeic preference pattern to be an aggregation of individual preferences, we do insist that group and individual preferences possess some fundamental coherence properties that are similar to those that are required for coherent epistemic preferences. These coherence properties are 1) to avoid sure exclusion, in that no individual be required to sacrifice its own welfare for the benefit of the group in every situation, and 2) that joint preference must not contradict individual conditional preferences. The first property ensures a weak sense of equanimity between individuals in the sense that no player is disadvantaged in all situations. The second property ensures that social relationships that exist between individuals are consistent with group social behavior.

The first property is the praxeic analogue to the epistemic notion of avoiding sure loss. To illustrate, let us consider the Prisoner’s Dilemma. Assume, in addition to the individual preference patterns $\{\succeq_i, \cong_i\}$, for $i \in \{x, y\}$ that there exists a group preference pattern $\{\succeq_{xy}, \cong_{xy}\}$, which orders all pairs $(u_x, u_y) \in \{C, D\} \times \{C, D\}$ such that $(u_x, u_y) \succeq_{xy} (u'_x, u'_y)$ means that the joint outcome (u_x, u_y) is at least as preferred for the group as the joint outcome (u'_x, u'_y) . Now suppose that $(C, C) \succeq_{xy} (D, C)$ and $(C, D) \succeq_{xy} (D, D)$; that is, it is always better for the group for X to cooperate, regardless of what Y does. This being the case, then it is reasonable that, in terms of X ’s own preference pattern, $C \succeq_x D$. In general, this requires, for an N -agent system, that

$$(u_1, \dots, u_i \dots u_N) \succeq_{xy} (u_1, \dots, u'_i \dots u_N) \forall u_j \in U_j, \\ \forall j \neq i \implies u_i \succeq_i u'_i.$$

This requirement prohibits situations where the good of the group always requires a player to sacrifice its own welfare (sure exclusion). Such situations are not self-enforcing, and hence would require either a dictator to command acquiescence or the player in question to be irrational.

To illustrate the second praxeic coherency requirement, consider the relation $C|C \succeq_x D|C$, which means that X would prefer to cooperate, given that Y would prefer to cooperate. Coherency then requires that $(C, C) \succeq_{xy} (D, C)$. In other words,

the individual preference for X to cooperate whenever Y prefers to cooperate is consistent with the group preference for both X and Y to cooperate. If this ordering did not hold, then the mere fact of X 's hypothetical assertion of Y preference would influence the preference of the group. In general, this coherency property requires, for an N -agent system, that

$$\begin{aligned} &u_i|(u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_N) \\ &\succeq_i u'_i|(u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_N) \\ \implies &(u_1, \dots, u_i, \dots, u_N) \\ &\succeq_{12\dots N} (u_1, \dots, u'_i, \dots, u_N). \end{aligned}$$

In other words, given X_j 's unconditional preferences for all $j \neq i$, then X_i 's conditional preference for u_i is sufficient to determine the group's preference for $(u_1, \dots, u_N)$.

Notice that these two praxeic coherency properties are exact analogues to the desired epistemic coherency properties. We can thus form a direct analogy between praxeic utility and epistemic utility, where the multiagent praxeic case is analogous to the multivariate epistemic case. The validity of this analogy will be established by demonstrating that the coherency requirements for the two contexts are satisfied by exactly the same mathematical structure.

C. A Logical Basis for Coherent Reasoning

In this section, we present three coherency desiderata and show how they motivate the construction of coherent utilities in both epistemic and praxeic contexts. The utilities that result will be shown to possess the mathematical structure of probability mass functions. We therefore establish that praxeology has a claim on the mathematics of probability theory that is just as valid as is the traditional epistemic claim. To facilitate this development, we will form praxeic analogues to the various epistemic concepts and present the epistemic and praxeic concepts in parallel.

We restrict to consideration of two-valued (i.e., Aristotelian) logic. In the epistemic context, this means that an event is either true or false. For any event set A , we say that A is true if any member of the set is true. The praxeic Aristotelian analogue to truth is instantiation—an action is either performed or it is not. For any set A of actions, we will say that A is instantiated if any member of the set is instantiated.

In the epistemic context, a basic notion to characterize an event in the quest for truth is plausibility, or the attractiveness of an option in terms of its veracity. In the praxeic context, events are characterized by their efficacy; that is, their capability to bring about the intended result of taking action.

1) *Desiderata for Coherent Evaluation:* In the epistemic context, we are interested in expressing degrees of plausibility regarding the truth of events under consideration, and in the praxeic context, we are interested in expressing degrees of efficacy regarding the instantiation of the actions under consideration. We first must generalize the concept of a utility. Conventionally, utilities are mappings from points in a sample space (in the epistemic context) or option space (in the praxeic context). Rather than restricting the utilities to be mappings

from points to real numbers, however, we wish to define the domain of the utilities as a Boolean algebra.

Jaynes [32] offers three basic desiderata that any theory of plausible (efficacious) reasoning ought to possess. The first desideratum is a standard one for all of utility-based decision theory.

D-1) Degrees of support are represented by real numbers.

All degrees of support (in both the epistemic and praxeic contexts) must be defined in the context of the environment that pertains to the problem. Let e denote the environment (i.e., the state of nature, which comprises all external factors that influence plausibility or efficacy). Since the plausibility (efficacy) of an event (option) may change as the environment changes, we will generally need to introduce additional notation to reflect this fact. Let $A \cdot e$ denote the event (option) A with respect to the environment e . We say that $A \cdot e$ is true (instantiated) if any element of A is true (instantiated) and e is the environment. We say that $AB \cdot e$ is true (instantiated) if both $A \cdot e$ and $B \cdot e$ are true (instantiated), and we say that $A|B \cdot e$ is true (instantiated) if $A \cdot e$ is true (instantiated) given that $B \cdot e$ is true (instantiated).

In accordance with Desideratum D-1, we must provide numerical values for preference orderings. Let us first consider the epistemic context and introduce the plausibility ordering $\{\succeq_P, \cong_P\}$ as defined in Section III-A. We may then define a utility function $f_P(A \cdot e)$ that satisfies (1), i.e., $A \cdot e \succeq_P B \cdot e$ if and only if $f_P(A \cdot e) \geq f_P(B \cdot e)$. In the praxeic context, we employ the ordering $\{\succeq_E, \cong_E\}$ meaning *{is as least as efficacious as, is as efficacious as}*. Let $f_E(A \cdot e)$ be a corresponding utility function denoting the numerical degree of efficacy of A being instantiated under e .

To ensure compatibility with conventional usage, we require the Boolean algebra to include all singleton sets. We will assume that the functions f_P and f_E possess a continuity property, such that an infinitesimal change in the degrees of plausibility (efficacy) will generate only infinitesimal changes in the numerical values that the corresponding utility functions assume. When our discussion applies to both utility functions, it will be convenient to suppress the subscript, and let f represent a utility function for either context.

We now apply the notion of *conditioning*, that is, the plausibility (efficacy) of one event (option), given that another event (option) is true (instantiated). Let A and B be arbitrary sets of events (options). Suppose that $B \cdot e$ is known to be true (instantiated). Then, $f(A|B \cdot e)$ denotes the conditional plausibility (efficacy) of $A \cdot e$ being true (instantiated) given that $B \cdot e$ is true (instantiated).

To motivate the second desideratum, suppose e is changed to a new environment e' in such a way that $f(A \cdot e') > f(A \cdot e)$ but that $f(B|A \cdot e') = f(B|A \cdot e)$. This should never decrease the plausibility (efficacy) of AB , that is, we require $f(AB \cdot e') \geq f(AB \cdot e)$. Furthermore, if a change in e makes A more plausible (efficacious), then it makes its complement less plausible (efficacious); i.e., if $f(A \cdot e') > f(A \cdot e)$, then $f(A^c \cdot e') < f(A^c \cdot e)$, where A^c is the complement of A . We summarize these requirements with the following desideratum.

D-2) Qualitative evaluations of support must agree with common sense.

D-3) If a conclusion can be obtained in more than one way while considering exactly the same issues, then every possible way must lead to the same result.

2) *Quantitative Rules of Behavior*: We now seek a consistent rule for relating the plausibility (efficacy) of AB to the plausibilities (efficacies) of A and B considered separately. We may evaluate the statement that A and B are both true (instantiated) under e by first considering $f(B \cdot e)$, the truth (instantiation) support for B under e , and then considering the statement that A is true (instantiated) given B under e , namely, $f(A|B \cdot e)$. Of course, since the intersection is commutative ($AB = BA$), we may reverse the roles of A and B without changing anything (this is in accordance with Desideratum D-3).

To elaborate, for a given environment, if both A and B are true (instantiated), then B must be true (instantiated). But if B is true (instantiated) then, for A also to be true (instantiated), it must be that A given B is true (instantiated). Furthermore, if B is false (not instantiated), then AB must also be false (not instantiated), regardless of whether or not A is true (instantiated). Thus, if we first consider the plausibility (efficacy) of B , then the plausibility (efficacy) of A will be relevant only if B is true (instantiated). Consequently, given the plausibilities (efficacies) of both B and $A|B$, we do not require the plausibility (efficacy) of A alone in order to compute the plausibility (efficacy) of AB .

If we reverse the roles of A and B , then we see that the plausibility (efficacy) of A and $B|A$ is also sufficient to determine the plausibility (efficacy) of BA . The upshot of this development is that the plausibility (efficacy) of AB is a function of the plausibilities (efficacies) of B and $A|B$ or, equivalently, of A and $B|A$. In other words,

$$f(AB \cdot e) = F[f(B \cdot e), f(A|B \cdot e)] = F[f(A \cdot e), f(B|A \cdot e)] \quad (2)$$

where F is some function to be determined.

Next, we observe that the consistency desideratum requires, when considering $f(ABC \cdot e)$, that if we consider BC first as a single event (option), application of (2) yields

$$f(ABC \cdot e) = F[f(BC \cdot e), f(A|BC \cdot e)]. \quad (3)$$

Now apply (2) to $f(BC \cdot e)$ to obtain

$$f(BC \cdot e) = F[f(C \cdot e), f(B|C \cdot e)]$$

which, when substituted into (3), yields

$$f(ABC \cdot e) = F\{F[f(C \cdot e), f(B|C \cdot e)], f(A|BC \cdot e)\}. \quad (4)$$

However, since set intersection is associative, we have $A(BC) = (AB)C$. Considering AB first, we also obtain

$$f(ABC \cdot e) = F[f(BC \cdot e), f(A|BC \cdot e)]$$

and repeated use of (2) also yields

$$f(ABC \cdot e) = F\{f(C \cdot e), F[f(B|C \cdot e), f(A|BC \cdot e)]\}. \quad (5)$$

The consistency desideratum requires that (4) and (5) must be the same. Thus, the function F must satisfy the following constraint, called the associativity equation [32], [33]:

$$F[F(x, y), z] = F[x, F(y, z)]. \quad (6)$$

A further constraint on F is that it also must satisfy the common-sense desideratum. To ensure this property, suppose e changes to e' such that $f(B \cdot e') > f(B \cdot e)$, but $f(A|B \cdot e') = f(A|B \cdot e)$. Then, common sense insists that $f(AB \cdot e') \geq f(AB \cdot e)$. In addition, we require that, if $f(B \cdot e') = f(B \cdot e)$ but $f(A|B \cdot e') > f(A|B \cdot e)$, then $f(AB \cdot e') \geq f(AB \cdot e)$. In other words, $F(x, y)$ must be nondecreasing in both arguments.

It remains to determine the structure of F that satisfies all of these constraints. By direct substitution, it is easily established that (6) is satisfied if

$$w[F(x, y)] = w(x) \cdot w(y) \quad (7)$$

for any function w . The following theorem establishes this as the general solution.

Theorem 1 (Cox, 1946): Suppose F is differentiable in both arguments, then (7) is the general solution to (6) for some positive, continuous, and monotonic function w . Consequently, for any events (options) A and B

$$w[f(AB \cdot e)] = w[f(A|B \cdot e)] \cdot w[f(B \cdot e)] \quad (8)$$

which is called the *product rule*, and

$$w[f(A \cdot e)] + w[f(A^c \cdot e)] = 1$$

which is called the *sum rule*. Furthermore

$$w[f(U \cdot e)] = 1$$

and

$$w[f(A \cup B \cdot e)] = w[f(A \cdot e)] + w[f(B \cdot e)] - w[f(AB \cdot e)].$$

For a proof of this theorem, see [34]–[36] or [32]. Jaynes observes that (6) was actually first solved by Abel as early as 1826 in a different context [37]. In addition, Aczél has established the same result without the assumption of differentiability [33].

3) *Constructing Utilities*: Let us now impose the additional assumption that w is nondecreasing. In addition, since w composed with f is a function of the events (options), we may, without loss of generality, define a function P over the events (options) as $P(A) = w[f(A \cdot e)]$ (since we now assume the environment is fixed, we may simplify the notation by dropping e from the argument list). Then, P possesses exactly the mathematical properties that are required to define probability over a Boolean algebra $\mathcal{B}$ of events (options), namely the following:

P-1: nonnegativity: $0 \leq P(A)$;

P-2: normalization: if U is the entire space, then $P(U) = 1$;

P-3: additivity: if A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$

for A, B in $\mathcal{B}$. These three properties are usually taken as axioms in standard expositions of probability theory as in Kolmogorov

[38]. These axioms are then used to derive all of the well-known probabilistic concepts, such as conditioning and independence. However, we see that Kolmogorov's axioms themselves are actually the consequences of more descriptive desiderata. In particular, conditioning, which is expressed merely as a definition with conventional treatments, is actually a fundamental concept under the more constructive approach offered by Cox. That is, (8) becomes the product rule of probability theory

$$P(AB) = P(A|B) \cdot P(B). \quad (9)$$

We have thus demonstrated that, with U a finite set of actions and given a preference pattern $\{\succeq_E, \cong_E\}$, there exists a mass function p_E , which characterizes the degree of support for instantiating the elements of U . To distinguish these special utilities from usual vN–M utilities, we will refer to utilities that satisfy the three desiderata (and, hence, also the axioms of probability theory) as *praxeic utilities*.

Just as multivariate probability mass functions are used to characterize multiple discrete random phenomena, we may use multivariate praxeic utility functions to characterize multiagent phenomena. These mass functions possess the desired coherency properties. To see, let U_1 and U_2 be two finite sets of singleton events (options), and let p_1 , p_2 , and p_{12} be the corresponding marginal and joint mass functions, respectively. If $p_{12}(u_1, u_2) \geq p_{12}(u'_1, u_2)$ for all $u_2 \in U_2$, then

$$p_1(u_1) = \sum_{u_2 \in U_2} p_{12}(u_1, u_2) \geq \sum_{u_2 \in U_2} p_{12}(u'_1, u_2) = p_1(u'_1)$$

which guarantees that sure loss and sure exclusion are avoided. Furthermore, if $p_{1|2}$ is a conditional mass function and if $p_{1|2}(u_1|u_2) \geq p_{1|2}(u'_1|u_2)$, then

$$\begin{aligned} p_{12}(u_1, u_2) &= p_{1|2}(u_1|u_2) \cdot p_2(u_2) \\ &\geq p_{1|2}(u'_1|u_2) \cdot p_2(u_2) = p_{12}(u'_1, u_2) \end{aligned}$$

which ensures that joint probability (efficacy) is consistent with conditional probability (efficacy).

IV. SATISFICING GAMES

A. Dual Utilities

The consequences of taking an action can be evaluated by two distinct desiderata of efficacy. The first deals with effectiveness, or how well the action achieves the fundamental purpose of the activity, and the second deals with efficiency, or how well the action conserves resources. The second desideratum exists because, in any practical situation, resources (money, energy, time, exposure to hazard, etc.) must be expended in order to accomplish whatever tasks are appropriate in the pursuit of the first desideratum. The vN–M approach is to aggregate the attributes that correspond to these concerns into a single preference pattern and a corresponding utility. For example, when buying an automobile, the attributes may be performance, style, purchase price, and operating costs. A vN–M utility would provide an ordering of the net effect of all of these attributes taken together.

However, total aggregation of attributes is not the only way to proceed. Suppose that we rank all options in terms of two preference patterns as follows. Let $\{\succeq_S, \cong_S\}$ denote binary ordering relationships meaning *{is at least as effective as, is as effective as}*, and let $\{\succeq_R, \cong_R\}$ denote binary ordering for *{is at least as inefficient as, is as inefficient as}*. It is useful to view a decision maker X as an entity with two personas, or roles. The *selecting persona*, denoted S , is X viewed exclusively in terms of effectiveness, that is, of achieving its goal without concern for the resources that may be expended. The *rejecting persona*, denoted R , is X viewed exclusively in terms of inefficiency, that is, of consuming resources without concern for achieving the goal. When viewed simultaneously from both perspectives, the agent is denoted as the concatenation of these to personas, i.e., $X = SR$. When more than one agent is under consideration, we will denote them as $X_1 = S_1R_1$ and $X_2 = S_2R_2$, etc.

We may associate a praxeic utility function with each persona. The utility p_S evaluates the options in terms of selectability, and p_R evaluates rejectability. Separating the attributes into two categories defines a dual-utilities approach, with which it is possible to define an alternative method of evaluating options. Instead of making global, or interoption, rankings by comparing each option to all others, one can make local, or intraoption, rankings for each option by comparing its effectiveness (achieving the goal) to its efficiency (consuming resources). Making such comparisons is consistent with human behavior. People commonly compare the pros versus the cons, the benefits versus the costs, the upside versus the downside. Making decisions based on local comparisons provides an alternative notion of rationality that is just as consistent with human nature as is the practice of making global comparisons. An option for which the selectability (effectiveness) is at least as great as its rejectability (inefficiency) will be termed a *satisficing* option. For the automobile-buying example, we may associate the attributes of performance and style with selectability and attributes purchase price and operating costs with rejectability. A vehicle for which the benefits (as determined by selectability) are at least as great as the costs (as determined by rejectability) is a satisficing choice. To put it in the vernacular, whereas an optimizer is committed to making the best, and only the best, choice, a satisficer is one who is content with getting his or her money's worth. We define an option to be *satisficingly rational* if its selectability equals or exceeds its rejectability.

B. Interdependence

In general, the interagent relationships of a complex society can be quite involved. An act by any individual member of a multiagent system has possible ramifications for the entire system. Some participants may be benefited by the act, some may be damaged, and some may be indifferent. Furthermore, although an individual may perform the act in its own interest or for the benefit of others or the entire society, the act is usually not implemented free of cost. Resources are expended, or risk is taken, or some other penalty or unpleasant consequence is incurred, perhaps by the individual whose act it is, perhaps by other participants and perhaps by the entire society. Although these costs may be defined independently from the benefits, the

measures associated with benefits and costs cannot be specified independently of each other due to the possibility of interaction. A critical aspect of modeling the behavior of an N -agent society, therefore, is the means of representing the interdependence of both positive and negative consequences of all possible joint options that could be undertaken. Since each decision maker has two personas, a global model of behavior must account for all $2N$ personas of the system.

Consider a set $\mathbf{X} = X_1 X_2 \cdots X_N$ of N agents and let U_i denote the (finite) option set for the i th agent. The set $\mathbf{S} = S_1 S_2 \cdots S_N$ is the collection of selecting personas, and $\mathbf{R} = R_1 R_2 \cdots R_N$ is the collection of rejecting personas. The *joint option set* is the product set $\mathbf{U} = U_1 \times U_2 \times \cdots \times U_N$, and the elements of $\mathbf{U}$ are vectors of the form $\mathbf{u} = (u_1, u_2, \dots, u_N)$, where $u_i \in U_i$. The *interdependence mass function* $p_{\mathbf{SR}}$, alternatively denoted $p_{S_1 \dots S_N R_1 \dots R_N}$, is a multivariate praxeic utility function; that is, $p_{\mathbf{SR}}(\mathbf{u}; \mathbf{v}) \geq 0 \forall (\mathbf{u}, \mathbf{v}) \in \mathbf{U} \times \mathbf{U}$ and $\sum_{\mathbf{u} \in \mathbf{U}} \sum_{\mathbf{v} \in \mathbf{U}} p_{\mathbf{SR}}(\mathbf{u}; \mathbf{v}) = 1$. The interdependence function characterizes all of the relationships that exist between all personas with respect to selecting option vector $\mathbf{u} = (u_1, \dots, u_N)$ and rejecting option vector $\mathbf{v} = (v_1, \dots, v_N)$. The special case when $v_i = u_i$ characterizes the conflict that arises between the two personas because of the desire to select an option on the basis of its effectiveness while desiring to reject it because of its lack of efficiency.

The *joint selectability and rejectability mass functions* may be obtained from the interdependence mass function as

$$p_{\mathbf{S}}(\mathbf{u}) = \sum_{\mathbf{v} \in \mathbf{U}} p_{\mathbf{SR}}(\mathbf{u}; \mathbf{v}) \quad (10)$$

$$p_{\mathbf{R}}(\mathbf{v}) = \sum_{\mathbf{u} \in \mathbf{U}} p_{\mathbf{SR}}(\mathbf{u}; \mathbf{v}). \quad (11)$$

The *individual selectability and rejectability marginals* may then be obtained as

$$p_{S_i}(u_i) = \sum_{\substack{u_j \in U_j \\ j=1, \dots, N \\ j \neq i}} p_{\mathbf{S}}(u_1, \dots, u_N) \quad (12)$$

$$p_{R_i}(u_i) = \sum_{\substack{u_j \in U_j \\ j=1, \dots, N \\ j \neq i}} p_{\mathbf{R}}(u_1, \dots, u_N). \quad (13)$$

C. Satisficing Decision Making

In Section III-B, we introduced the notion of making choices on the basis of getting one's money's worth. We now make that concept mathematically precise. By virtue of restricting the selectability and rejectability utilities to be mass functions, a decision maker has a unit of mass to allocate among the options to characterize their selectability, and it has a unit of mass to allocate to characterize the rejectability. Let $\mathbf{X} = X_1 X_2 \cdots X_N$ be a group of N agents with joint option space $\mathbf{U}$. A *satisficing game* [39]–[43] is a triple $(\mathbf{X}, \mathbf{U}, p_{\mathbf{SR}})$, where $p_{\mathbf{SR}}$ is an interdependence function. The *jointly satisficing solution at caution level q* of a satisficing game is the subset of all option vectors

such that the joint selectability is at least as great as the caution level times the joint rejectability, that is

$$\Sigma_q = \{\mathbf{u} \in \mathbf{U} : p_{\mathbf{S}}(\mathbf{u}) \geq q p_{\mathbf{R}}(\mathbf{u})\}$$

where $p_{\mathbf{S}}$ and $p_{\mathbf{R}}$ are the joint selectability and rejectability functions obtained from (10) and (11), respectively.

A *satisficing decision at caution level q* for an individual X_i is any element of the set

$$\Sigma_q^i = \{u \in U_i : p_{S_i}(u) \geq q p_{R_i}(u)\} \quad (14)$$

where p_{S_i} and p_{R_i} are the i th marginals of $p_{\mathbf{S}}$ and $p_{\mathbf{R}}$, respectively. The parameter q is called the *index of caution* and reflects the decision maker's emphasis. This quantity may be viewed as a "tuning" parameter to regulate the relative weight that the decision maker wishes to attribute to the goal-seeking and resource-seeking attributes of the options. Nominally, we set $q = 1$, thereby attributing equal weight to the two categories of attributes. As is discussed in [44], this quantity can be used to facilitate negotiations between decision makers.

The solution sets for a satisficing game are generally not singleton sets. The set Σ_q consists of all joint options that are deemed to be good for the group, and the sets Σ_q^i consist of all individual options that are good enough for the individual. The relationships between these sets is rather complex and is discussed in [44], where it is shown that every element of an individual's satisficing set is an element of some jointly satisficing set. This result is a consequence of avoiding sure exclusion and provides a basis for meaningful negotiation.

D. Social Interaction

To illustrate the difference between satisficing theory and conventional theory, we contrast it with vN–M game theory. A vN–M strategic game is a triple $(\mathbf{X}, \mathbf{U}, \mathbf{f})$, where $\mathbf{f} = (f_1, \dots, f_N)$ is a set of vN–M utilities that define the payoffs for each player as a function of the joint option vectors (i.e., $f_i(\mathbf{u})$ is X_i 's payoff if $\mathbf{u}$ is jointly instantiated). The distinction between this game and a satisficing game is that the interdependence function is not a vN–M utility, since it is not a function of the actions that the players may take. Rather, it is a function of the preferences for action.

The interdependence function characterizes all of the relationships that exist among the members of the group. Because of its structure as a multivariate mass function, we may use the properties of independence and conditioning to facilitate its construction. Thus, two personas are *independent* if their joint praxeic utility function factors into the product of the corresponding marginals. For example, if the selectability persona of X_i and the selectability persona of X_j have no influence on each other, then $p_{S_i S_j}(u_i, u_j) = p_{S_i}(u_i) \cdot p_{S_j}(u_j)$ for all $(u_i, u_j) \in U_i \times U_j$. If the selectability persona of X_i and the rejectability persona of X_k have no influence on each other, then $p_{S_i R_k}(u_i, u_k) = p_{S_i}(u_i) \cdot p_{R_k}(u_k)$ for all $(u_i, u_k) \in U_i \times U_k$, and so forth.

In addition, one persona is *conditioned* on another persona if the preferences that the first persona assumes are

influenced by the preferences that the second persona assumes. Thus, $p_{S_i|S_j}(u_i|u_j)$ is a conditional mass function that characterizes the selectability that X_i ascribes to u_i , given that agent X_j prefers to select u_j . By the product rule, the joint selectability of the personas S_i and S_j is $p_{S_i S_j}(u_i, u_j) = p_{S_i|S_j}(u_i|u_j) \cdot p_{S_j}(u_j)$. Similarly, $p_{S_i|R_k}(u_i|u_k)$ is a conditional mass function that characterizes the selectability that X_i ascribes to u_i , given that agent X_k prefers to reject u_k . In addition, the interdependence of personas S_i and R_k is $p_{S_i R_k}(u_i, u_k) = p_{S_i|R_k}(u_i|u_k) \cdot p_{R_k}(u_k)$. Notice that social preferences are deontologically neutral; they may model selfishness, benevolence, malevolence, or even indifference. Thus, social utilities permit the agents to extend their spheres of interest beyond the self.

The product rule may be generalized to form the chain rule, which constitutes a factorization of the interdependence into products of conditional and marginal mass functions. Let $\pi = \{S_1 S_2 \dots S_N R_1 R_2, \dots, R_N\}$ denote the collection of all personas in an N -agent system, and for any subset $\pi_i \subset \pi$, let π_i^c denote the complement of π_i . With this notation, the interdependence mass function is denoted p_π . Now let $\{\pi_1, \pi_2, \dots, \pi_M\}$, $M \leq 2N$, be a partition of π , that is, $\pi_i \cap \pi_j = \emptyset$ for $i \neq j$ and $\bigcup_{i=1}^M \pi_i = \pi$. Repeated applications of the product rule yields the chain rule (with arguments suppressed)

$$p_\pi = p_{\pi_1|\pi_1^c} \cdot p_{\pi_2|\pi_1^c \pi_2^c} \cdot p_{\pi_3|\pi_1^c \pi_2^c \pi_3^c} \cdots p_{\pi_{M-1}|\pi_1^c \dots \pi_{M-1}^c} \cdot p_{\pi_M}.$$

For example, if $N = 3$, let $\pi_1 = S_1 S_2$, $\pi_2 = S_3 R_2$, and $\pi_3 = R_1 R_3$. Then

$$p_{S_1 S_2 S_3 R_1 R_2 R_3} = p_{S_1 S_2|S_3 R_1 R_2 R_3} \cdot p_{S_3 R_2|R_1 R_3} \cdot p_{R_1 R_3}.$$

If any of the personas are independent of other personas, then the chain rule may be simplified. For example, if S_1 and S_2 are independent of R_1 and R_3 , S_3 and R_2 are independent of R_3 , and R_1 is independent of R_3 ; then, we have

$$p_{S_1 S_2 S_3 R_1 R_2 R_3} = p_{S_1 S_2|S_3 R_2} \cdot p_{S_3 R_2|R_1} \cdot p_{R_1} \cdot p_{R_3}. \quad (15)$$

In [44], a number of examples are provided to illustrate the structure and function of social utilities.

V. CONCLUSION

With the satisficing approach, the relationship between individual and group interests is neither top-down nor bottom-up. Rather, we may characterize the relationship as inside-out, in the sense that both individual and group preferences emerge as consequences of the way the conditional preferences propagate through the system via the chain rule. In other words, neither the individuals nor the group form their preferences in a social vacuum; rather, they form them as *a posteriori* results of the social relationships that naturally exist between the participants. This phenomenon of emergence seems to comply with the observation of the philosopher Ortega y Gasset, who noted that “Order is not pressure which is imposed on society from without, but an equilibrium which is set up from within” [45].

By avoiding the necessity of *a priori* imposing either individual or group preference, the contradictions between group and individual interests that are inherent with individual rationality are mitigated, if not completely resolved. This is the fundamental difference between conventional game theory, which uses asocial utilities, and hence can only accommodate egoism, and satisficing game theory, which uses social utilities, and hence can accommodate sophisticated social behaviors such as cooperation, compromise, negotiation, and altruism, as well as egoism.

By eschewing optimization as the ideal and replacing it with a precise notion of satisficing (i.e., being good enough), we remove the wedge that optimization places between group and individual interests. Comparing optimizing with satisficing reveals an interesting tradeoff between the extent of interest and the way preferences are compared. Optimal behavior is strictly an interest-local concept—only individuals can optimize—and compliance with that paradigm can be achieved with asocial utilities by making global interooption comparisons (comparing the attributes of one option to the same attributes of other options). By contrast, cooperative behavior is generally an interest-global enterprise—only groups can cooperate—and compliance with that paradigm can be achieved with dual social utilities by making local intraoption comparisons (comparing different attributes of the same option). Thus, individual rationality and satisficing rationality offer a rather interesting parallel between localization and globalization. Under the former, global (i.e., interooption) information is used to define local (individual) interest, while under the latter, local (i.e., intraoption) information is used to define global (group) interest.

Extending the sphere of interest beyond the self increases the complexity of a multiagent system model, since it must account for sophisticated social relationships. As noted by Palmer, however, “Complexity is no argument against a theoretical approach if the complexity arises not out of the theory itself but out of the material which any theory ought to handle” [46]. While satisficing game theory can be more complex than the vN–M approach, it possesses the following desirable attributes:

- 1) it is parsimonious in that it is not more complex than it needs to be to function in an ecologically sound way;
- 2) it is internally consistent in that contradictory requirements are not imposed on the agents;
- 3) it is *comprehensive* in that it accounts for all relevant relationships that exist among the individual agents.

Social utilities and satisficing game theory provide a new tool for the designer of artificial multiagent systems. It supplements, rather than supplants, existing approaches, such as vN–M game theory, that are founded on individual rationality. It is not a panacea that provides a universal framework for all multiagent systems. For applications where individual rationality provides the appropriate framework, satisficing theory may not provide significantly different results. However, for situations where cooperation is of paramount importance, social utilities and satisficing game theory may provide a viable and tractable approach.

REFERENCES

- [1] H. Raiffa, *Decision Analysis*. Reading, MA: Addison-Wesley, 1968.

- [2] K. J. Arrow, "Economic theory and the hypothesis of rationality," in *Utility and Probability*, J. Eatwell, M. Milgate, and P. Newman, Eds. New York: Norton, 1990, pp. 25–37.
- [3] R. Nozick, *The Nature of Rationality*. Princeton, NJ: Princeton Univ. Press, 1993.
- [4] J. Harsanyi, "Morality and the theory of rational behavior," *Social Res.*, vol. 44, no. 4, 1977.
- [5] A. Tversky and D. Kahneman, "Rational choice and the framing of decisions," in *Rational Choice*, R. M. Hogarth and M. W. Reder, Eds. Chicago, IL: Univ. Chicago Press, 1986.
- [6] G. Weiss, Ed., *Multiagent Systems*. Cambridge, MA: MIT Press, 1999.
- [7] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 2003.
- [8] A. Rubinstein, *Modeling Bounded Rationality*. Cambridge, MA: MIT Press, 1998.
- [9] T. J. Sargent, *Bounded Rationality in Macroeconomics*. Oxford, U.K.: Oxford Univ. Press, 1993.
- [10] G. Lilly, "Bounded rationality," *J. Economic Dynamics Control*, vol. 18, pp. 205–230, 1994.
- [11] T. Sandholm and V. Lesser, "Coalitions among computationally bounded agents," *Artif. Intell.*, vol. 94, no. 1, pp. 99–137, 1997.
- [12] E. Sober and D. S. Wilson, *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard Univ. Press, 1998.
- [13] A. Bergson, "A reformulation of certain aspects of welfare economics," *Quarterly J. Economics*, vol. 52, pp. 310–334, 1938.
- [14] P. A. Samuelson, *Foundations of Economic Analysis*. Cambridge, MA: Harvard Univ. Press, 1948.
- [15] P. Monaghan, "Taking on 'rational man'," *Chronicle Higher Educ.*, Jan. 24, 2003.
- [16] L. A. Zadeh, "What is optimal?," *IRE Trans. Inform. Theory*, vol. 4, no. 1, p. 3, 1958.
- [17] J. von Neumann and O. Morgenstern, *The Theory of Games and Economic Behavior*. Princeton, NJ: Princeton Univ. Press, 1944.
- [18] R. Axelrod, *The Evolution of Cooperation*. New York: Basic Books, 1984.
- [19] W. Poundstone, *Prisoner's Dilemma*. New York: Doubleday, 1992.
- [20] R. D. Luce and H. Raiffa, *Games and Decisions*. New York: Wiley, 1957.
- [21] M. Shubik, *Game Theory in the Social Sciences*. Cambridge, MA: MIT Press, 1982.
- [22] K. J. Arrow, *Social Choice and Individual Values*, 2nd ed. New York: Wiley, 1951.
- [23] T. S. Ferguson, *Mathematical Statistics*. New York: Academic, 1967.
- [24] D. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions*. New York: Dover, 1979.
- [25] L. J. Savage, *The Foundations of Statistics*. New York: Dover, 1972.
- [26] M. DeGroot, *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.
- [27] J. Harsanyi, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge, U.K.: Cambridge Univ. Press, 1977.
- [28] A. K. Sen, *Collective Choice and Social Welfare*. Amsterdam, The Netherlands: North-Holland, 1979.
- [29] M. Taylor, *The Possibility of Cooperation*. Cambridge, U.K.: Cambridge Univ. Press, 1987.
- [30] A. K. Sen, "Rational fools: A critique of the behavioral foundations of economic theory," in *Beyond Self-Interest*, J. J. Mansbridge, Ed. Chicago, IL: Univ. Chicago Press, 1990, ch. 2.
- [31] *The Compact Oxford English Dictionary*. Oxford, U.K.: Oxford Univ. Press, 1991.
- [32] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
- [33] J. Aczél, *Lectures on Functional Equations and Their Applications*. New York: Academic, 1966.
- [34] R. T. Cox, "Probability, frequency, and reasonable expectation," *Amer. J. Phys.*, vol. 14, pp. 1–13, 1946.
- [35] ———, *The Algebra of Probable Inference*. Baltimore, MD: The Johns Hopkins Univ. Press, 1961.
- [36] M. Tribus, *Rational Descriptions, Decisions and Designs*. New York: Pergamon, 1969.
- [37] N. H. Abel, *Oeuvres Complètes de Niels Henrik Abel*, L. Sylow and S. Lie, Eds. Christiania, Denmark: Impr. de Groendahl & søen, 1881.
- [38] A. N. Kolmogorov, *Foundations of the Theory of Probability*, 2nd ed. New York: Chelsea, 1956.
- [39] W. C. Stirling and M. A. Goodrich, "Satisficing games," *Inf. Sci.*, vol. 114, pp. 255–280, Mar. 1999.
- [40] W. C. Stirling, M. A. Goodrich, and D. J. Packard, "Satisficing equilibria: A nonclassical approach to games and decisions," *Autonomous Agents Multi-Agent Systems*, vol. 5, pp. 305–328, 2002.
- [41] W. C. Stirling, "Games machines play," *Minds Machines*, vol. 12, pp. 327–352, 2002.
- [42] ———, "Beyond optimization: Overcoming the limitations of individual rationality," in *Game Theoretic Decision Theoretic Agents (GTDT'02): Papers From the AAI Workshop Tech. Rep. WS-02-06*, Edmonton, AB, Canada, 2002, pp. 99–108.
- [43] ———, *Satisficing Games and Decision Making: With Applications to Engineering and Computer Science*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
- [44] W. C. Stirling and R. L. Frost, "Social utility functions, Part 2: Applications," *IEEE Trans. Syst., Man, Cybern., C: Appl. Rev.*, vol. 35, no. 4, pp. 533–543, Nov. 2005.
- [45] J. Orgeta y Gasset, "Mirabeau: An Essay on the Nature of Statesmanship," Historical Conservation Society, Manila, 1975.
- [46] F. R. Palmer, *Grammar*. Harmondsworth, Middlesex: Penguin, 1971.

Wynn C. Stirling received the B.A. Hons. degree *magna cum laude* in mathematics and the M.S. degree in electrical engineering from the University of Utah, Salt Lake City, in 1969 and 1971, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1983.

From 1972 to 1975, he was with Rockwell International Corporation, Anaheim, CA, and from 1975 to 1984 he was employed by ESL, Inc., Sunnyvale, CA, where he was responsible for the development of multivehicle trajectory reconstruction capabilities.

He joined the faculty of the Department of Electrical and Computer Engineering at Brigham Young University, Provo, UT, in 1984, where he is a Professor. He is the author or coauthor of more than 70 publications. He is coauthor of a graduate-level text titled *Mathematical Methods and Algorithms for Signal Processing* (Englewood Cliffs, NJ: Prentice-Hall, 2000) and is the author of a monograph titled *Satisficing Games and Decision Making: With Applications to Engineering and Computer Science* (Cambridge, U.K.: Cambridge Univ. Press, 2003). His current research interests include multiagent decision theory, estimation theory, information theory, and stochastic processes.

Dr. Stirling is a Member of Phi Beta Kappa and Tau Beta Pi. He has served on the program committees for conferences on imprecise probability theory and multiagent decision theory.