Int J Soc Robot (2009) 1: 53–69
DOI 10.1007/s12369-008-0007-x

ORIGINAL PAPER

Satisficing Coordination and Social Welfare for Robotic Societies

Wynn C. Stirling · Matthew S. Nokleby

Accepted: 28 October 2008 / Published online: 26 November 2008

Abstract The design of robotic systems that are capable of sophisticated social behavior such as cooperation, compromise, negotiation, and altruism, requires more complex mathematical models than are afforded by the classical mechanisms for making value judgments and decisions. A new concept of multi-agent satisficing, defined in terms of relative effectiveness and efficiency, is an alternative to classical optimization-based decision making. Conditional utilities, which take into account the interests of others as well as the self, represent an alternative to the categorical utilities of classical decision theory. A multi-agent utility aggregation structure is developed that avoids the sure subjugation of the interests of any individual to the interests of the group. By expressing a society as a directed acyclic graph, Bayesian network theory is applied to artificial societies. A satisficing social welfare function accounts for the influence relationships among decision-making agents.

Keywords Multi-agent systems · Game theory · Social choice theory · Satisficing · Conditional preferences · Coherence

1 Introduction

Multi-stakeholder decision problems arise in many contexts, including social choice theory, game theory, distributed control theory, multi-criterion/multi-objective decision theory, multi-agent systems theory, and social robotics. Although the particulars of these various contexts can differ widely, to be rational, each must possess two fundamental attributes: (a) an ability to make value judgments regarding alternatives, and (b) procedures for using value judgments to make choices.

The field of social robotics, in particular, provides a rich environment for the application of decision-making logics that are able to accommodate sophisticated social behaviors such as compromise, negotiation, and altruism. Whether a social robot interfaces with humans, other robots, or both, it typically resides in a community that involves some notion of coordination (which may be either cooperative or competitive). In such an environment, value judgments can depend upon the desires and preferences of others, and procedures for making choices must take these complex social relationships into account.

The two approaches that dominate conventional multi-agent decision theory are noncooperative game theory and social choice theory. Noncooperative game theory focuses primarily on the welfare of the individuals without consideration of the welfare of the group, and is designed to characterize decision making in an environment of competition and market driven expectations. Social choice theory, on the other hand, focuses on the welfare of the group, and there is no guarantee that any individual's preferences will be accommodated by the group-level decision. Thus, both conventional approaches place a wedge between individual and group interests.

The main contribution of this paper is the presentation of an approach to multiagent decision making that removes, or at least mitigates, this wedge. We provide a mathematical framework within which to design and synthesize complex decision-making collectives that are able to accommodate socially complex decision making. This framework permits

W.C. Stirling (✉) · M.S. Nokleby
The Electrical and Computer Engineering Department, Brigham Young University, Provo, UT 84602, USA
e-mail: wynn@ee.byu.edu

M.S. Nokleby
e-mail: nokleby@ee.byu.edu

group and individual preferences to be reconciled in a way that respects both points of view and does not categorically disadvantage one over the other. There are three key elements of this framework. First, we define conditional utilities that permit agents to modulate their preferences over action profiles as a function of the preferences of others as well as of the self; this approach results in a true generalization of the classical game-theoretic utility structure. We also show how this formulation relates to classical solution concepts such as Nash equilibrium and the Nash bargaining solution. Second, we impose a condition of coherence that prohibits the unconditional subjugation of any individual to the will of the group. Third, we replace the conventional emphasis on optimization with a more socially accommodating concept of satisficing, or being good enough. This change in emphasis permits compromise solutions to be considered that are simultaneously acceptable to all individuals and to the group as a whole.

Section 2 provides a brief history of classical multi-agent decision making and motivates our approach. Section 3 introduces the key components of the framework we are proposing, Sect. 4 introduces new concepts of social welfare, Sect. 5 reconciles our theory with classical approaches, Sect. 6 describes a special case of what we term *decoupled* social systems, and Sect. 7 offers conclusions.

2 Background

Cooperative robotics is an active research area. Of particular interest is the development of theories for decentralized control of multi-robot societies. Swarm-based approaches have demonstrated the emergence of cooperative behavior [23, 24]. Potential functions, consisting of constraints and goals that are imposed upon the system, have been used to address the mobile robot navigation problem [4]. Shannon information theory has been applied to the investigation of diversity among heterogeneous agents, thereby enabling an assessment of the ability of the system to perform cooperatively [3]. Behavior-based approaches have been applied to the design of cooperative robotic teams, stressing minimalism, statelessness, and tolerance [52]. The variety displayed by these various of approaches is a strong indication of the complexity involved in the design of cooperative multiagent systems, and is in indication that there is no single approach that can be universally applied to the design and synthesis of such systems.

Because of the complexity of multiagent systems, it is important to review the fundamental principles that are exploited, either implicitly or explicitly, in their design. Accordingly, we provide a brief review of classical decision-theoretic foundations and a discussion of rationality.

2.1 Classical Decision-Theoretic Foundations

The multi-stakeholder decision problem originated in the social sciences context, with foundations laid by Bergson [5], Samuelson [37], Arrow [1], and others, who assert that individual values are the fundamental elements of a society. Arrow has provided what is perhaps the most clear definition of this concept: “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 standard he deems relevant” [1, p. 17]. Furthermore, Friedman argues that the process by which these preferences are obtained is irrelevant: “The economist has little to say about the formation of wants; this is the province of the psychologist. The economist’s task is to trace the consequences of any given set of wants. The legitimacy of any justification for this abstraction must rest ultimately, in this case as with any other abstraction, on the light that is shed and the power to predict that is yielded by the abstraction” [16, p. 13]. According to the Arrow/Friedman model, each participant in a multi-agent decision problem comes to the decision-making activity with pre-defined preference orderings, the origins of which are not germane to the decision problem. Such preference orderings are *categorical*. The assumption that each individual possesses a categorical preference ordering has been adopted almost universally in classical multi-stakeholder decision-making contexts.

The most common procedure for using value judgments to make choices is to invoke some notion of optimization—the *sine qua non* of classical decision theory. What is optimal in multi-stakeholder settings, however, can depend upon the point of view. In the classical game-theoretic context, each individual seeks to optimize value to itself, and a Nash equilibrium is a constrained mutually optimal solution for all players in the sense that no individual can unilaterally improve its welfare by changing its decision. On the other hand, in the social choice context, it is the “organization incarnate,” as Raiffa put it [35], who seeks to maximize value for the group considered as a whole. In the former case, the value of the individual decisions to the group is not explicitly considered, and in the latter case, although the value judgments of the individuals are used to define group-level decisions (e.g., a weighted sum of individual valuations), there is no assurance that the resulting decision will maximize the value to any individual. In fact, the decision that is best for the group can be extremely unfavorable to some members of the group.

2.2 Rationality

The classical approach to decision making in group settings is the doctrine of individual rationality: the notion that each individual should act in a way that maximizes its own satisfaction (without explicit regard for the satisfaction of others). This doctrine enjoys a central role in classical decision theory and game theory. As discussed by Tversky and Kahneman, "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" [51, p. 89].

The uncritical application of individual rationality as a model for decision making in multi-agent contexts can be problematic. Arrow has observed that "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" [2, p. 203].

If all agents are indeed focused on, and only on, their narrow self-interest, then categorical preferences are appropriate. Difficulties arise, however, when the sphere of concern of an individual extends beyond its own narrow self-interest. The only way such an individual can use categorical preferences to accommodate the preferences of other individuals is to redefine its values by substituting (at least partially) the values of the others for its own. Such behavior is a manifestation of *categorical altruism*, i.e., irrevocably sacrificing one's own welfare in an attempt to benefit another, thus fundamentally changing the nature of the association.

Considerable research, notably in the field of behavorial economics [7], has addressed the need for agents to define their preferences such that they consider social interactions. Fehr and Schmidt [14] discuss how individual preference orderings may be modified to take into account concepts such as fairness and cooperation by introducing a notion of inequity aversion. To account for this attribute, they include, in addition to a purely selfish component, an inequity aversion component in their utility. Consequently, they rely upon (re-defined) categorical preference orderings to model social interactions. All that changes is the definition of the individual's self-interest. This approach, however, has serious limitations, as acknowledged by Sen: "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" [38, p. 19]. Categorical altruism simulates cooperation, compromise, and altruism with a regime that is explicitly designed to characterize selfishness, competition, and avarice, and does not offer a natural and intuitively pleasing framework within which to express sophisticated and complex social relationships. While such constructions may serve to explain some forms of human behavior, it is difficult to see how they can be used systematically to synthesize complex relationships between artificial agents.

The foundational assumptions of categorical preferences for each individual and optimization (either for individuals or for the group) undergird virtually all of classical formalized decision theory in both individual and group settings. These assumptions correspond to *analysis* tools that serve, with varying success, to explain and predict human behavior, but they are not causal: they do not govern human behavior. On the other hand, models that are used to design a system of artificial autonomous decision-making agents must be causal: they are *synthesis* tools that will indeed govern the behavior of the artificial society.

Many social science researchers argue, however, that the classical foundational assumptions do not provide an adequate model for human behavior (e.g., see [25, 45]). And if their adequacy to analyze human behavior is questioned, then we may rightly question their appropriateness as assumptions with which to synthesize the behavior of artificial societies that are expected to behave in ways that are can be understood and trusted by humans. As Shubik as acknowledged, "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 [41]."

3 A Social Framework for Cooperative Decision Making

A social welfare function, as defined by Arrow, is "a process or rule which, for each set of individual orderings R_1, R_2, ..., R_n for alternative social states (one ordering for each individual), states a corresponding social ordering of alternative social states R" [1, p. 23]. Classically, the individual orderings (either ordinal or cardinal), are categorical, in that they account only for the interests of the individuals. We wish to expand the spheres of interest of the individuals to include the interests of others as itself. However, once we move beyond restricting to individual interests, the notion of optimization becomes problematic. Optimization is an individual concept: for a group to optimize, it must act as a single unit, capable of making rational judgments and choices. Such a structure however, is not consistent with our assumption that the decision-makers are autonomous.

Our approach is to replace the twin assumptions of optimization and categorical preferences with two alternative concepts: satisficing and conditioning. Our goal is to create a satisficing social welfare function and individual welfare functions that can be used to construct compromise solutions that are simultaneously acceptable to the group and the individuals, thereby removing, or at least reducing, the wedge that separates classical concepts of group and individual interests.

3.1 Satisficing

In a multi-agent setting it is not generally possible to maximize both individual and group preferences simultaneously. A potentially more socially accommodating concept is that decisions are "good enough." What is best for you may be different from what is best for me, but what is good enough for you may also be good enough for me, provided we have some flexibility in our respective notions of what it means to be good enough. The term "satisficing" has been advanced as a synonym for this alternative to strict optimization.

The first usage of the term "satisficing" in a decision-theoretic context is attributed to Simon [42–44], who addressed the question of how a decision maker might make a choice in the presence of informational or computational limitations. Simon's approach is to seek an optimal choice, but to terminate searching and once the decision maker's aspiration level has been met. Put another way, to satisfice is to accept the best solution so far obtained, once the cost of continuing to search exceeds the expected improvement in value were the search to continue. Many other variations of this concept have appeared in the literature [6, 13, 19, 22, 27, 28, 32, 34, 39, 49, 50, 53–55], and it is not the intent of this paper to review them in detail. Suffice it to say, however, that all of these approaches view satisficing as a species of bounded rationality: one settles for a solution that is deemed to be "good enough," but which is not necessarily, and usually not, optimal in any meaningful sense. Satisficing *à la* Simon is an heuristic approximation to the ideal of being best (and is only constrained from achieving this ideal by practical limitations).

The concept of satisficing presented herein differs from the afore-mentioned notion in several important ways.[1] First, in contrast to satisficing as advanced by Simon and others, it is not heuristic; rather, it provides a concept of satisficing that is as mathematically formalized and precise as is the notion of optimization. Second, it treats the notion of being good enough as the ideal (rather than an approximation)—it is *not* a species of bounded rationality. Third, it extends to the multi-agent case, thereby providing a natural framework for multi-agent decision making. Fourth, it readily accommodates the extension of interests beyond the self, thereby accommodating more sophisticated social relationships than self-interest affords. We retain the term "satisfice" because, even though our approach is not heuristic, we nevertheless seek solutions that are good enough, with the essential difference being that we provide a non-heuristic definition of what it means to be good enough.

Although it seems eminently reasonable at least to attempt (given sufficient resources) to seek an optimal decision, humans often invoke a systematic approach to decision making (even in single-agent decision problems) that, while still based on quantitative measures of performance, does not correspond to optimization. In the vernacular, the optimization paradigm corresponds to seeking "the best and only the best" solution. Also common, however, is the paradigm of "getting your money's worth." In an intuitively pleasing sense, this latter notion admits an interpretation as being good enough, and it is this concept that we invoke as the satisficing paradigm that we develop in this paper. A comprehensive introduction to this perspective can be found in [47].

Many theorists (e.g., [1, 12, 17, 26]) have argued that it is unwise to aggregate conflicting interests into a single preference ordering. Some have asserted that in a social setting individuals have multiple facets, as defined by Steedman and Krause [46], who maintain that an agent, although an indivisible unit, nevertheless is capable of considering its choices from different points of view, and that separate utilities may be defined to correspond to each facet of an individual. A natural way to classify attributes is according to their effectiveness and efficiency. Each individual may be viewed as being composed of two facets: the *selecting facet*, which evaluates actions in terms of effectiveness toward pursuing objectives without concern for efficiency, and the *rejecting facet*, who evaluates actions in terms of efficiency with respect to consuming resources without concern for effectiveness. We shall view these selecting and rejecting facets as the "atoms" of the society,

When formulating a problem under the satisficing framework, it is essential that the selecting and rejecting criteria not be restatements of each other. The selecting criterion should correspond to the goals of the problem, and the rejecting criterion should correspond to the consumption of resources. This dual utilities approach is the basis for our notion of satisficing.

[1] Some of the material in this section has been discussed in other venues (e.g., [47, 48]), but is included here as a tutorial to ensure a self-contained development.

Under the optimization paradigm, all of the performance measures are combined into a single utility, whereas under the satisficing paradigm, the measures of effectiveness are encoded separately from the measures of efficiency. Under the optimization paradigm, the alternatives are compared against each other in order to identify the globally best one. By contrast, under the satisficing paradigm, the effectiveness and efficiency attributes are locally compared for each alternative separately, and all alternatives for which the effectiveness measures exceed the efficiency measures are considered to be satisficing. Thus, whereas the optimization paradigm is designed to identify a single best alternative, the satisficing paradigm is designed to identify all alternatives that are good enough. The non-uniqueness attribute is a key feature of satisficing in a multi-stakeholder environment, since it is amenable to flexibility on the part of the individuals and of the group.

To introduce the formalism of satisficing, let us first consider a single agent X, with selecting and rejecting facets denoted S and R, respectively, and let u_S denote the selecting utility, or *selectability*, which measures the progress toward the goal of X, and u_R denotes the rejecting inutility, or *rejectability*, which measures the consumption of resources such as cost, exposure to hazard, loss of social reputation, and so forth.

Definition 1 Let $\mathcal{A}$ denote the set of actions available to X. An action $a \in \mathcal{A}$ is *satisficing* if $u_S(a) \geq qu_R(a)$ where $q \in [0, 1]$ regulates the threshold for rejecting elements of $\mathcal{A}$ as not satisficing. The *satisficing set* is

$$\Sigma_q = \{a \in \mathcal{A}\colon u_S(a) - qu_R(a) \geq 0\}. \tag{1}$$

The parameter q represents X's attitude regarding the weight it affords the different facets of the decision problem. Nominally, $q = 1$, indicating equal weight apportioned to the selecting and rejecting criteria. As we shall see, however, q can serve as a measure of how willing the individual is to accommodate the interests of the group.

Satisficing as defined above is expressed in a single-agent context with categorical utilities. It is easily seen, in this simple context, that u_S and u_R can easily be combined to form a classical utility $u_X(a) = u_S(a) - qu_R(a)$, which is amenable to optimization. Optimization, however, is designed to produce a single best solution, whereas, by contrast, satisficing is designed to produce a (possibly) non-singleton set of solutions that are good enough in the sense that the effectiveness of the action is as least as great as its inefficiency. In the single-agent context, satisficing represents a novel approach, but if it is possible to optimize, then there may be little incentive to seek a satisficing solution. The real power of the satisficing concept, however, is manifest in the multi-agent case, as will be further developed below.

3.2 Conditioning

Let $\mathbf{X} = (X_1, \dots, X_n)$ denote a collective of autonomous stake-holders (e.g., agents). More specifically, let $\mathbf{S} = (S_1, \dots, S_n)$ denote the collective of selecting facets, and let $\mathbf{R} = (R_1, \dots, R_n)$ denote the collective of rejecting facets. Notationally, we write $\mathbf{V} = \mathbf{SR} = (S_1, \dots, S_n, R_1, \dots, R_n)$, a system of $2n$ facets. Since we will be dealing with the facets, rather than the agents, it is convenient to use the symbol V_i, $i = 1, \dots, 2n$, to denote either a selecting facet or a rejecting facet.

Let $\mathcal{A}_i$ denote a finite set of alternatives available to X_i. Of course, if X_i takes action $a_i \in \mathcal{A}_i$, then that action also applies to S_i and R_i (split personalities are not allowed, but this does not mean that S_i and R_i must always contemplate taking the same action). The product action space is denoted $\boldsymbol{\mathcal{A}} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$, and an *action profile* $\mathbf{a} = (a_1, \dots, a_n) \in \boldsymbol{\mathcal{A}}$ denotes the joint action taken by the collective.

A *categorical utility* for V_i, denoted u_{V_i}, is a mapping $u_{V_i}\colon \boldsymbol{\mathcal{A}} \to \mathbb{R}$, and provides a total ordering of all action profiles for V_i. According to the conventional Arrow/Friedman model, categorical utilities for all participants in the multi-stakeholder decision problem are defined prior to the decision-making activity and, furthermore, the mechanisms that dictate the way they are defined are irrelevant. As an alternative, we introduce the notion of a *conditional utility*. To develop this concept, we must first define a *commitment*.

Definition 2 Let V_i be an arbitrary element of $\mathbf{V}$, and let $\mathbf{V}_j = (V_{j_1}, \dots, V_{j_k})$ be an arbitrary k-element subset of $\mathbf{V}$ that does not include V_i. A *commitment profile* $\{\mathbf{a}_{j_1}, \dots, \mathbf{a}_{j_k}\}$, $\mathbf{a}_{j_\ell} \in \boldsymbol{\mathcal{A}}$, is a hypothetical statement by V_i that the action profile $\mathbf{a}_{j_\ell}$ is the one that is to which V_{j_ℓ} is constrained, $\ell = 1, \dots, k$.

Definition 3 A *conditional utility* for V_i with respect to a commitment profile $\{\mathbf{a}_{j_1}, \dots, \mathbf{a}_{j_k}\}$, denoted $u_{V_i|\mathbf{V}_j}(\mathbf{a}|\mathbf{a}_{j_1}, \dots, \mathbf{a}_{j_k})$, is a utility for V_i given that $\mathbf{V}_j$ is committed to $\{\mathbf{a}_{j_1}, \dots, \mathbf{a}_{j_k}\}$.

This definition is a generalization of the original concept of conditional utilities in that the conditioning is with respect to the entire commitment profile of each X_{j_ℓ}, $l = 1, \dots, k$, rather than with respect to the individual commitments of each X_{j_ℓ} (see Sect. 6). The resulting utility structure represent a true generalization of traditional categorical utilities and enables us to reconnect our theory with classical game theory.

Operationally, a conditional utility for V_i serves as the consequent of a hypothetical proposition whose antecedent is a commitment by $\mathbf{V}_j$. This expression does not represent V_i's actual utility of $\mathbf{a}$, nor does it imply that V_{j_ℓ} is constrained to $\mathbf{a}_{j_\ell}$. Instead, it means that, if $(V_{j_1}, \dots, V_{j_k})$ were simultaneously to be constrained to $\{\mathbf{a}_{j_1}, \dots, \mathbf{a}_{j_k}\}$, then V_i would define its utility of $\mathbf{a}$ accordingly.

An attractive feature of a conditional utility is that it permits V_i to express *conditional altruism*. To illustrate, suppose $u_{V_j}(\mathbf{a}) \gg u_{V_j}(\mathbf{a}')$, that is, V_j were to ascribe much higher categorical utility to $\mathbf{a}$ than to $\mathbf{a}'$, but V_i were to do the opposite, ascribing higher utility to $\mathbf{a}'$ than to $\mathbf{a}$, i.e., $u_{V_i}(\mathbf{a}') \geq u_{V_i}(\mathbf{a})$. V_i could give deference to V_j by replacing its categorical utility u_{V_i} with a conditional utility $u_{V_i|V_j}$ such that $u_{V_i|V_j}(\mathbf{a}|\mathbf{a}) \geq u_{V_i|V_j}(\mathbf{a}'|\mathbf{a})$ but $u_{V_i|V_j}(\mathbf{a}'|\mathbf{a}') = u_{V_i|V_j}(\mathbf{a}|\mathbf{a}') = u_{V_i}(\mathbf{a}')$, thus deferring to V_j if, but only if, V_j were to favor $\mathbf{a}$ strongly over $\mathbf{a}'$.

3.3 Social Networks

Conditional preferences provide each individual with the ability to define its preferences as a function of the hypothetical preferences of all other subsets of the collective. This feature represents an important departure from the traditional categorical definitions of preference and provides the foundation for the modeling of a complex society that possesses sophisticated social relationships such as altruism (either benevolent or malevolent). Conditional preference relations permit the explicit modeling of such relationships, rather than merely simulating them by redefining categorical preference orderings. Although conditional preference relations are more complex than are categorical ones, as noted by Palmer, "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" [31, p. 176].

Nevertheless, the introduction of conditional utilities increases the complexity of the mathematical model of a collective. At one extreme, all of the members of the collective would be devoted to narrow self-interest, and all utilities would be categorical (the classical game-theoretic model). At the other extreme, each of the members would be influenced by the preferences of every other member, resulting in a fully connected collective. Fortunately, however, many potentially interesting societies are such that the connections between the members are relatively sparse. Just as with human societies, it is likely that members will be organized into relatively small clusters of individuals that are somewhat loosely connected with other clusters. One such model is a hierarchical structure, where the preferences of superiors influence those of subordinates. Another, more parallel model, is one where the individuals are grouped into function, spatial, or temporal neighborhoods. A powerful and convenient way to represent such relationships is through graph theory, which provides a means to express directly the influence relationships that exist among the individuals. With such a formalism, the vertices of the graph represent the members of the collective, and the edges represent the influence flows among them as encoded in the conditional utilities. For the extreme case where all individuals possess categorical preferences, the graph would have no edges—each individual would be expressed by an isolated vertex. When conditional preferences exist, however, the graph will have edges, as illustrated in Fig. 1.

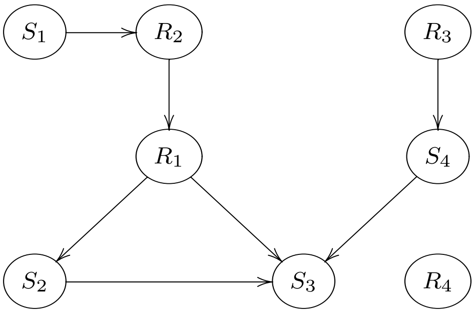


Fig. 1 A directed acyclic graph

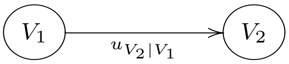
In this paper we concentrate on *directed acyclic graphs*, or DAGs. Formally, a directed graph is a pair $\mathcal{G} = (\mathbf{V}, E)$, where $\mathbf{V} = (V_1, \dots, V_{2n})$ is a finite set of vertices and E is a set of *edges* linking pairs of vertices. If V_j is directly influenced by V_i but V_j does not directly influence V_i, then there is a directed edge, denoted "$\rightarrow$" from V_i to V_j. A *path* from V_i to V_j is a sequence of vertices $\{V_i, V_{k_1}, V_{k_2}, \dots, V_j\}$ such that $V_i \rightarrow V_{k_1} \rightarrow V_{k_2} \rightarrow \cdots \rightarrow V_j$. We write $V_i \mapsto V_j$ if there is a path from V_i to V_j. If there are no paths such that $V_i \mapsto V_i$ for any i, the graph is said to be *acyclic*.

If $V_i \rightarrow V_j$, then V_i is called a *parent* of V_j, and V_j is a *child* of V_i. The set of parents of V_i is denoted $\mathrm{pa}\,(V_i) = \{V_{i_j}\colon V_{i_j} \rightarrow V_i, j = 1, \dots, p_i\}$, and the set of *children* of V_i is denoted $\mathrm{ch}\,(V_i)$. If V_i has no parents, then $\mathrm{pa}\,(V_i) = \varnothing$. The descendents of V_i, denoted $\mathrm{de}\,(V_i)$, is the subset of vertices $\{V_{i_m}\colon V_i \mapsto V_{i_m},\ m = 1, \dots, d_i\}$.

Let $\mathrm{cp}\,(V_i) = \{\mathbf{a}_{i_1}, \dots, \mathbf{a}_{i_{p_i}}\}$ denote the commitment profile for $\mathrm{pa}\,(V_i)$. For each V_i, $u_{V_i|\mathrm{pa}\,(V_i)}[\mathbf{a}|\,\mathrm{cp}\,(V_i)]$ is the utility that V_i ascribes to $\mathbf{a}$, given that V_{i_j} commits to $\mathbf{a}_{i_j}$, $j = 1, \dots, p_i$. If V_i has no parents, the conditional utility is the categorical utility; i.e., $u_{V_i|\mathrm{pa}\,(V_i)} = u_{V_i}$ if $\mathrm{pa}\,(V_i) = \varnothing$. Consider the DAG illustrated in Fig. 1. By inspection, $\mathrm{pa}\,(S_1) = \mathrm{pa}\,(R_3) = \mathrm{pa}\,(R_4) = \varnothing$, $\mathrm{pa}\,(S_2) = \{R_1\}$, $\mathrm{pa}\,(S_3) = \{S_2, S_4, R_1\}$, $\mathrm{pa}\,(S_4) = \{R_3\}$, $\mathrm{pa}\,(R_1) = \{R_2\}$, and $\mathrm{pa}\,(R_2) = \{S_1\}$.

A fundamental property of a DAG is the *Markov condition*: nondescendent nonparents of a vertex have no influence on the vertex, given the state of its parent vertices [9]. Consequently, if a society can be represented as a DAG, the conditional utility of a facet is dependent only upon the commitments of its parents. Thus, for the DAG in Fig. 1, R_2 is influenced only by the commitments of S_1, S_3 is influenced by the commitments of S_2, S_4, and R_1, and so forth. Thus, conditional utility of R_2 is of the form $u_{R_2|\mathrm{cp}\,(R_2)}$,

Fig. 2 A two-agent DAG



where $\mathrm{cp}\,(R_2) = \{S_1\}$, and the conditional utility of S_3 is of the form $u_{S_3|\mathrm{cp}\,(S_3)}$, where $\mathrm{cp}\,(S_3) = \{S_2, S_4, R_1\}$. Categorical utilities are associated with the root vertices, S_2, R_3, and R_4, since these vertices have no parents.

4 Social Welfare

4.1 Collective Preferences

The central question for a collective of autonomous decision makers is how they should function as a group. In the classical non-cooperative game-theoretic formulation, the notion of a group preference is irrelevant—each individual is committed to, and only to, its own satisfaction, and the emergence of a coherent notion of group welfare would be strictly coincidental. As observed by Shubik, "It may be meaningful, in a given setting, to say that group 'chooses' or 'decides' something. It is rather less likely to be meaningful to say that the group 'wants' or 'prefers' something" [40, p. 124]. Social choice theory, on the other hand, focuses on the aggregation of individual preferences to form a social welfare function that can be used to define what is best for the group. Classical social choice theory, however, as developed by Arrow [1], Debreu [11], Fishburn [15], and others, also relies upon categorical preferences, as does multi-objective decision theory [20]. The main classical result, attributed to Debreu, is that a necessary and sufficient for a group utility to be defined as the weighted sum of individual utilities is that the individual utilities must be categorical.

In the presence of conditional preferences, the issue of social welfare takes on added complexity. For example, the traditional axioms of social choice theory, such as the independence of irrelevant alternatives, becomes problematic. Thus, we must pursue a different course when aggregating conditional preferences. In the interest of clarity, we begin our discussion of this concept with the bi-facet case, with $\mathbf{V} = (V_1, V_2)$. Let us suppose that V_1 possesses a categorical utility u_{V_1} and V_2 possesses a conditional utility $u_{V_2|V_1}$. The corresponding DAG is displayed in Fig. 2. Given these utilities, the central questions are: (i) Can these two utilities be combined in a rational way to form a group utility? and, if so, (ii) How should they be combined?

To address this issue, we introduce the notion of a joint commitment. A *joint commitment* by (V_1, V_2) is a condition that, simultaneously, V_1 is committed to $(a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$ and V_2 is committed to $(a_1', a_2') \in \mathcal{A}_1 \times \mathcal{A}_2$. The utility of a joint commitment would provide a complete description of the way the collective views all possible consequence profiles (one for each facet). It would provide information regarding the degree of conflict and the possibilities for compromise, since only one profile can actually be implemented by the collective.

If there were no conflicts, then there would exist a joint commitment of the form $[(a_1^*, a_2^*), (a_1^*, a_2^*)]$ that would simultaneously maximize benefit to V_1 and V_2 and, hence, by the Pareto principle, to the collective. In the presence of conflicts, however, joint commitments of the form $[(a_1, a_2)(a_1, a_2)]$, where both commit to the same profile, would represent a compromise solution. The issue, then, is to define an acceptable compromise.

To determine the utility of a joint commitment, consider the following argument. For $[(a_1, a_2), (a_1', a_2')]$ to be a joint commitment, it is necessary that (a_1, a_2) be a commitment by V_1. But if (a_1, a_2) is a commitment by V_1, then for (a_1', a_2') to be a commitment by V_2, (a_1', a_2') must be a commitment by V_2 given that (a_1, a_2) is a commitment by V_1. Furthermore, if (a_1, a_2) is not a commitment by V_1, then $[(a_1, a_2), (a_1', a_2')]$ is not a joint commitment, regardless of whether or not (a_1', a_2') is a commitment by V_2. Thus, when considering the utility of a joint commitment to $[(a_1, a_2), (a_1', a_2')]$, if the utility of a commitment to (a_1, a_2) by V_1 is considered first, then the utility of a commitment to (a_1', a_2') by V_2 will be relevant only if (a_1, a_2) is a commitment by V_1. Consequently, given the utility of a commitment to (a_1, a_2) by V_1 and the conditional utility of a commitment to (a_1', a_2') by V_2 given that (a_1, a_2) is a commitment by V_1, knowledge of the categorical utility of a commitment to (a_1', a_2') by V_2 is not required in order to compute the utility of a joint commitment to $[(a_1, x_a), (a_1', a_2')]$. Thus, the utility of a joint commitment to $[(a_1, a_2), (a_1', a_2')]$ is a function of the categorical utility of a commitment to (a_1, a_2) by V_1 and the conditional utility of a commitment by V_2 to (a_1', a_2') given that (a_1, a_2) is a commitment by V_1.

Let $\hat{u}_{V_1V_2}$ denote the *utility of a joint commitment*. By the above arguments, this function can be expressed as

$$\begin{aligned}&\hat{u}_{V_1V_2}[(a_1, a_2), (a_1', a_2')]\\ &\quad = F[u_{V_1}(a_1, a_2), u_{V_2|V_1}(a_1', a_2'|a_1, a_2)] \qquad (2)\end{aligned}$$

for some function F, called the *aggregation function*.

4.2 Aggregation

Obviously, there are many possibilities for F, and to narrow the choices, it is necessary to impose some additional constraints. One reasonable constraint is that the collective possess at least a weak sense of equity so that a meaningful notion of cooperation can occur. Specifically, we wish to avoid a condition of *categorical subjugation*. To introduce this concept, let us restrict interest to the collective S_1

and V_2. We shall say that S_1 is categorically subjugated to the collective if every consequence profile that is acceptable to the collective would require S_1 to sacrifice its performance. Suppose that

$$u_{S_1}(a_1', a_2') > u_{S_1}(a_1'', a_2''), \tag{3}$$

but

$$\hat{u}_{S_1 V_2}[(a_1', a_2'), (a_1, a_2)] < \hat{u}_{S_1 V_2}[(a_1'', a_2''), (a_1, a_2)] \tag{4}$$

for all $(a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$. Then S_1 would be categorically subjugated, since S_1's preferred joint action can never be preferred by the society. Avoiding categorical subjugation ensures that all participants have a "seat at the table" when negotiating. Otherwise, the interests of some facets will be so contrary to the interests of the collective that, no matter what the collective decides, the interests of the affected individual facets will be suppressed. Unless the possibility (although not the guarantee) exists that the interests of the individual are compatible with the interests of the collective, the individual will be effectively disenfranchised. Although categorical subjugation is not always avoided in human societies (e.g., dictatorships), avoiding categorical subjugation is an important feature of an artificial society that must negotiate to reach a compromise.

If categorical subjugation is to be avoided, then there must exist an action profile $(\tilde{a}_1, \tilde{a}_2)$ such that, if (3) holds, then

$$\hat{u}_{S_1 V_2}[(a_1', a_2'), (\tilde{a}_1, \tilde{a}_2)] \geq \hat{u}_{S_1 V_2}[(a_1'', a_2''), (\tilde{a}_1, \tilde{a}_2)]. \tag{5}$$

A similar argument regarding the categorical subjugation of, say, R_1 can be made with the inequalities reversed in (3), (4), and (5) when R_1 replaces S_1.

The question now becomes: what conditions are necessary to impose upon the aggregation function F to ensure that categorical subjugation can never occur? To address this question, let us turn to an analogous issue. A Dutch book is a gambling situation such that, no matter what the outcome, the gambler will be worse off for having taken the gamble—a situation of *sure loss* (one's reward is always less than one's stake). To illustrate a Dutch book, Suppose Y can take one of two distinct values: y_1 or y_2, and let $q(y)$ denote a belief function of y; that is $q(y)$ measures the strength of belief that $Y = y$. Without loss of generality, we may restrict belief functions to the unit interval; that is, $0 \leq q(y) \leq 1$. (We refrain from using the term "probability" here, since we do not require q to possess all of the properties of a probability mass function.)

By convention, we will assume that we have full belief that exactly one of these values obtains, that is, that the disjunction of y_1 and y_2 must occur, and that beliefs are additive, thus,

$$q(y_1 \vee y_2) = q(y_1) + q(y_2) = 1. \tag{6}$$

Now let Z take on one of two distinct values z_1 or z_2, and let $r(z, y)$ denote the belief that $Z = z$ and $Y = y$ simultaneously. Let us now assume that

$$q(y_2) > q(y_1) \tag{7}$$

$$r(z_1, y_2) < r(z_1, y_1) \tag{8}$$

$$r(z_2, y_2) < r(z_2, y_1). \tag{9}$$

The following example illustrates a Dutch book. Suppose you purchase a \$1 gamble that $Y = y_2$, and deem a fair purchase price to be $q(y_2)$; that is, you pay \$$q(y_2)$ for the gamble to win \$1. Now also suppose you sell the gamble $(z_1, y_2) \vee (z_2, y_2)$. By additivity of beliefs, a fair selling price for this bet would be $r[(z_1, y_2) \vee (z_2, y_2)] = r(z_1, y_2) + r(z_2, y_2)$. However, according to the above ordering, you must have $q(y_2) > \frac{1}{2}$ and, since $r(z_1, y_2) + r(z_2, y_2) < r(z_1, y_1) + r(z_2, y_1)$, it follows that $r[(z_1, y_2) \vee (z_2, y_2)] < \frac{1}{2}$. After all gambles have been bought and sold, your net wealth is $r[(z_1, y_2) \vee (z_2, y_2)] - q(y_2) < 0$. To overcome this loss, you hope to make up the difference once the outcome of the gamble is known. But if neither y_2 nor $(z_1, y_2) \vee (z_2, y_2)$ occur, you win nothing and you pay nothing, and if $(z_1, y_2) \vee (z_2, y_2)$ occurs, then, of course, y_2 occurs, so you win \$1 which you must pay to the buyer of your gamble. Thus, once the gambles have been bought and sold, your net wealth is invariant to whatever happens—you suffer a sure loss.

A belief system is said to be *coherent* if it is not possible to construct a Dutch book. The Dutch Book Theorem [10, 36] and its converse [21] state that a belief system is coherent if and only if it complies with a probability measure that describes the degrees of belief regarding the propositions under consideration. The above example does not comply with the laws of probability theory, since $q(y_2) \neq r(z_1, y_2) + r(z_2, y_2)$; that is, marginalization fails.

The above discussion illustrates the fact that categorical subjugation and sure loss are mathematically equivalent. Thus, if a multi-agent valuation system is to be coherent, in that it is not possible to construct a situation where categorical subjugation can occur, then the valuation system must comply with the mathematical structure of probability theory.

Definition 4 Let u_{V_i} denote a categorical utility for V_i. The collective $\mathbf{V}$ is *coherent* if, for each $i \in \{1, \dots, 2n\}$, given that $u_{V_i}(\mathbf{a}) > u_{V_i}(\mathbf{a}')$, there exists a commitment sub-profile $(\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_{i-1}, \tilde{\mathbf{a}}_{i+1}, \dots, \tilde{\mathbf{a}}_{2n})$ such that

$$\hat{u}_{V_1 \cdots V_{2n}}(\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_{i-1}, \mathbf{a}, \tilde{\mathbf{a}}_{i+1}, \dots, \tilde{\mathbf{a}}_{2n}) \\ \geq \hat{u}_{V_1 \cdots V_{2n}}(\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_{i-1}, \mathbf{a}', \tilde{\mathbf{a}}_{i+1}, \dots, \tilde{\mathbf{a}}_{2n}) \tag{10}$$

if V_i is a selecting facet, with the inequalities reversed if V_i is a rejecting facet.

Let $\mathbf{V}$ be a group of decision making facets whose influence relationships can be expressed with a directed acyclic graph. For each V_i, let $\mathrm{pa}\,(V_i) = (V_{i_1}, \ldots, V_{i_{p_i}})$ denote the p_i parents of V_i, and let $\boldsymbol{\mathcal{A}}^{p_i} = \boldsymbol{\mathcal{A}} \times \cdots \times \boldsymbol{\mathcal{A}}$ ($p_i$times) denote the p_i-fold product of the joint action space corresponding to the parents of V_i. If V_i has no parents, then $\boldsymbol{\mathcal{A}}^{p_i} = \varnothing$. Let $\mathrm{cp}\,(V_i) = (\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_{p_i}})$ denote the commitment profile for $\mathrm{pa}\,(V_i)$. For each V_i, $u_{V_i|\mathrm{pa}\,(V_i)}[\mathbf{a}|\,\mathrm{cp}\,(V_i)]$ is the utility that V_i ascribes to $\mathbf{a}$, given that V_{i_j} commits to $\mathbf{a}_{i_j}$, $j = 1, \ldots, p_i$. If V_i has no parents, the conditional utility is the categorical utility; i.e., $u_{V_i|\mathrm{pa}\,(V_i)} = u_{V_i}$ if $\mathrm{pa}\,(V_i) = \varnothing$.

Theorem 1 *If a society can be represented as a directed acyclic graph, categorical subjugation cannot occur if and only if the utilities $u_{V_i|\mathrm{pa}\,(V_i)}$ are conditional mass functions. That is,*

$$u_{V_i|\mathrm{pa}\,(V_i)}[\mathbf{a}|\,\mathrm{cp}\,(V_i)] \geq 0 \quad \forall \mathbf{x} \in \boldsymbol{\mathcal{A}} \tag{11}$$

and

$$\sum_{\mathbf{a}\in\boldsymbol{\mathcal{A}}} u_{V_i|\mathrm{pa}\,(V_i)}[\mathbf{a}|\,\mathrm{cp}\,(X_i)] = 1 \tag{12}$$

for all $(\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_{p_i}}) \in \boldsymbol{\mathcal{A}}^{p_i}$. *Furthermore, the utility of a joint commitment to* $(\mathbf{a}_1, \ldots, \mathbf{a}_{2n})$ *is*

$$\hat{u}_{\boldsymbol{V}}(\mathbf{a}_1, \ldots, \mathbf{a}_{2n}) = \prod_{i=1}^{2n} u_{V_i|\mathrm{pa}\,(V_i)}[\mathbf{a}_i\,|\,\mathrm{cp}\,(V_i)] \tag{13}$$

or, more specifically,

$$\hat{u}_{\boldsymbol{SR}}(\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{a}'_1, \ldots, \mathbf{a}'_n) = \prod_{i=1}^{n}\prod_{j=1}^{n} u_{S_i|\mathrm{pa}\,(S_i)}[\mathbf{a}_i\,|\,\mathrm{cp}\,(S_i)]u_{R_j|\mathrm{pa}\,(R_j)}[\mathbf{a}'_j\,|\,\mathrm{cp}\,(R_j)], \tag{14}$$

where $\mathbf{a}_i$ is the commitment by S_i, $i = 1, \ldots, n$ and $\mathbf{a}'_j$ is the commitment by R_j, $j = 1, \ldots, n$.

Proof Mathematically (albeit with different semantics), we may view V_i as random variables defined over the sample spaces $\mathcal{V}_i$, $i = 1, \ldots, 2n$. The Dutch Book Theorem and its converse establish that the necessary and sufficient condition to ensure that sure loss (categorical subjugation) cannot occur is that $u_{V_i|\mathrm{pa}\,(V_i)}$ must correspond to the conditional probability mass functions of V_i given $\mathrm{cp}\,(V_i)$. Thus, the categorical utilities of the root vertices must possesses the mathematical structure of marginal probability mass function and the conditional utility of non-root vertices possesses the mathematical structure of conditional probability mass functions. Consequently, the vertices and edges of the DAG satisfy all of the conditions of a Bayesian network, and we may apply the fundamental theorem of Bayesian networks; namely, that the joint probability mass function of the random variables associated with the vertices is the product of the conditional probability mass functions of all non-root vertices, and the marginal mass functions of all root vertices [8, 18, 33]. Equation (13) is simply an application of the law of compound probability. Thus, coherence is established. □

We will term utilities that comply with Theorem 1 *praxeic utilities*. It should be noted that this formulation requires all utilities to be non-negative and sum to unity. This restriction, however, does not reduce the generality of the theorem, since utilities can be subjected to positive affine transformations without affecting the solution.

Equation (14) expresses the values of the selecting and rejecting facets simultaneously. Since parents of selecting facets may comprise both selecting and rejecting facets and similarly for the parents of rejecting facets, this function contains all of the possibilities for compromise and conflict. To be useful for decision making, however, it is necessary to compute the joint selectability for all joint commitments by the selecting facets, and the joint rejectability for all joint commitments by the rejecting facets. Since $\hat{u}_{\boldsymbol{SR}}$ is a multivariate mass function, we may compute the *joint selectability and rejectability marginals* as

$$\hat{u}_{S}(\mathbf{a}_1, \ldots, \mathbf{a}_n) = \sum_{\mathbf{a}'_1,\ldots,\mathbf{a}'_n} \hat{u}_{\boldsymbol{SR}}(\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{a}'_1, \ldots, \mathbf{a}'_n) \tag{15}$$

$$\hat{u}_{R}(\mathbf{a}'_1, \ldots, \mathbf{a}'_n) = \sum_{\mathbf{a}_1,\ldots,\mathbf{a}_n} \hat{u}_{\boldsymbol{SR}}(\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{a}'_1, \ldots, \mathbf{a}'_n). \tag{16}$$

Once the joint selectability and rejectability marginals have been computed, we are in a position to define a satisficing social welfare function. We first observe that, since only one action profile can be implemented, to make a decision, we must ascribe the same commitment to each facet, yielding the *joint praxeic selectability* and *joint praxeic rejectability*

$$\tilde{u}_{\boldsymbol{S}}(\mathbf{a}) = \hat{u}_{\boldsymbol{S}}(\mathbf{a}, \ldots, \mathbf{a}) \tag{17}$$

and

$$\tilde{u}_{\boldsymbol{R}}(\mathbf{a}) = \hat{u}_{\boldsymbol{R}}(\mathbf{a}, \ldots, \mathbf{a}). \tag{18}$$

We next define *satisficing social welfare function*

$$W(\mathbf{a}) = \tilde{u}_{\boldsymbol{S}}(\mathbf{a}) - q_G\tilde{u}_{\boldsymbol{R}}(\mathbf{a}) \tag{19}$$

and the *jointly satisficing set*

$$\boldsymbol{\Sigma}_{q_G} = \{\mathbf{a}\colon W(\mathbf{a}) \geq 0\}. \tag{20}$$

The parameter q_G is a measure of the relative weight the group as a whole places on the selecting criteria versus the rejecting criteria. Nominally, $q_G = 1$, indicating equal weight. If, however, the context of the problem is to place high weight on individual performance, it may be appropriate to reduce q_G, which will increase the cardinality of the jointly satisficing set and make it easier to reach a compromise (as will be discussed further below). All consequence profiles for which the satisficing social welfare is non-negative are deemed to be satisficing for the group.

We may also compute the *individual selectability and rejectability marginal utilities* as

$$\tilde{u}_{S_i}(\mathbf{a}_i) = \sum_{\neg \mathbf{a}_i} \hat{u}_{\boldsymbol{S}}(\mathbf{a}_1, \ldots, \mathbf{a}_n) \tag{21}$$

and

$$\tilde{u}_{R_i}(\mathbf{a}_i) = \sum_{\neg \mathbf{a}_i} \hat{u}_{\boldsymbol{R}}(\mathbf{a}_1, \ldots, \mathbf{a}_n), \tag{22}$$

where we have employed the so-called *not-sum* notation; namely, $\sum_{\neg \mathbf{a}_i}$ to mean that the sum is taken over all $\mathbf{a}_j$ for $j \neq i$.

The individual welfare function is

$$W_i(\mathbf{a}) = \tilde{u}_{S_i}(\mathbf{a}_i) - q_i \tilde{u}_{R_i}(\mathbf{a}_i) \tag{23}$$

and the *individually satisficing set* is

$$\Sigma^i_{q_i} = \{\mathbf{a}\colon W_i(\mathbf{a}) \geq 0\}, \tag{24}$$

where q_i is X_i's relative weight regarding the selecting and rejecting criteria. This parameter is X_i's *index of negotiation*. Setting $q_i < 1$ attributes more weight to the selecting criteria than the rejecting criteria, and enlarges the individual satisficing set, thereby reducing X_i's threshold for an acceptable outcome.

A satisficing set (either for the group or individuals) constitutes the set of consequences for which effectiveness, as measured by the selectability utility, is at least as great as q_i times the inefficiency, as measured by the rejectability inutility. Rather than focusing on seeking the best and only the best solution, the satisficing methodology focuses on eliminating bad solutions. Since the satisficing set eliminates all alternatives whose effectiveness does not exceed their efficiency, it possesses a weak notion of optimality; namely, the maximum number of unacceptable choices are eliminated. If, in the extreme case, all but one choice are eliminated, then the satisficing solution coincides with the optimal solution.

The *compromise set* is the set of all joint actions that are simultaneously satisficing for the group and for the individuals; that is,

$$\mathcal{C} = \boldsymbol{\Sigma}_{q_G} \cap \Sigma^1_{q_1} \cap \cdots \cap \Sigma^n_{q_n}. \tag{25}$$

In practice, the q_i's indicate the attitudes of the individuals with respect to how much they value effectiveness versus inefficiency. As will be discussed in Sect. 6, they can be interpreted as negotiation parameters that can be adjusted in a way that indicates the individuals' willingness to modify their valuation systems in an attempt to reach a compromise solution.

Table 1 The payoff matrix for the conventional Prisoner's Dilemma game

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

Example 1 (The Social Prisoner's Dilemma) The conventional Prisoner's Dilemma game is designed to characterize behavior between two decision makers in an environment were cooperation leads to better results than does defection but, if only one attempts to cooperate, that individual becomes vulnerable to being exploited by the other. Classically, this game is defined in terms of categorical utilities. Let C and D denote cooperation and defection, respectively. The corresponding categorical utilities are the entries of the payoff matrix displayed in Table 1. The joint option (C, C) (next best for both) is the Pareto optimal solution, while (D, D) (next worst for both) is the Nash equilibrium solution. Notice that the game is symmetrical. The classical assumption for this game is that there is no social relationship between the players, and that each is intent on, and only on, maximizing its own welfare, regardless of the effect doing so has on the other.

Now let us add some social context to this problem. Suppose a leader-follower relationship exists between them, with X_1 being the leader and X_2 the follower. We shall assume that X_1 follows the conventional structure of maximizing payoff, but X_2 is interested in (a) following the lead of X_1, (b) resisting exploitation, and (c) not offending X_1 by taking advantage of the possible propensity for X_1 to cooperate. Notice that this structure results in an asymmetrical relationship between the players. We shall take the definition of selectability as the same as with the conventional formulation; namely, to seek to maximize payoff. For rejectability, however, we invoke a component that is not present in the conventional formulation; namely, to account for social issues, and assume that the players have a unit of social resource they may commit to each outcome.

Since the leader has no social commitments, we take rejectability for X_1 to be the same for each outcome. Accordingly, the categorical selectability and rejectability values for the leader are provided in Table 2.

To account for the social context, we take the utilities for X_2 to be conditional, and assume that both selectability and

Table 2 Selectability and rejectability utilities for X_1 for the social Prisoner's Dilemma game

	(C, C)	(C, D)	(D, C)	(D, D)
u_{S_1}	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{4}{10}$	$\frac{2}{10}$
u_{R_1}	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

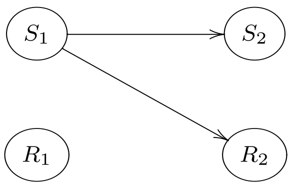


Fig. 3 The influence network for the social Prisoner's Dilemma game

Table 3 S_2's conditional selectability for the social Prisoner's Dilemma game

(x_1, x_2)				
	(C, C)	(C, D)	(D, C)	(D, D)
$u_{S_2\|S_1}(x_1, x_2\|C, C)$	1	0	0	0
$u_{S_2\|S_1}(x_1, x_2\|C, D)$	1	0	0	0
$u_{S_2\|S_1}(x_1, x_2\|D, C)$	1	0	0	0
$u_{S_2\|S_1}(x_1, x_2\|D, D)$	0	0	0	1

the rejectability of X_2 are influenced by the selectability of X_1, as indicated in Fig. 3.

Table 3 displays the conditional selectability for X_2 given the selection commitments of X_1. If X_1 were to commit to selecting either (C, C), (C, D), or (D, C), then X_2, in the interest of attempting to cooperate, would place all of its conditional selectability on (C, C) (mutual cooperation). But if X_1 were to commit to selecting (D, D), then X_2 would place all of its conditional selectability on (D, D) (mutual defection).

Table 4 displays the conditional rejectability for X_2 given the commitments of X_1. If X_1 were to commit to selecting (C, C), then X_2 would place zero conditional rejectability on that outcome and apportion the conditional rejectability equally among the other outcomes. If X_1 were to commit to (C, D), X_2 would place all of its rejectability on that outcome to ensure it will not be exploited. If X_1 were to commit to (D, C), then X_2 would not reject that outcome so as not to exploit X_1, and instead would reject exploitation by placing its conditional rejectability on (C, D). Finally, if X_1 were to commit to (D, D), X_2 would not reject that outcome, but would instead reject (C, D) as before.

The utility of a joint commitment is given by

$$\begin{aligned} &u_{S_1 S_2 R_1 R_2}[(x_1, x_2), (x'_1, x'_2), (y_1, y_2), (y'_1, y'_2)] \\ &\quad = u_{S_1}(x_1, x_2) u_{S_2|S_1}(x'_1, x'_2)|x_1, x_2) \\ &\qquad \times u_{R_1}(y_1, y_2) u_{R_1|S_1}[(y'_1, y'_2)]. \end{aligned} \tag{26}$$

Table 4 R_2's conditional rejectability for the social Prisoner's Dilemma game

(x_1, x_2)				
	(C, C)	(C, D)	(D, D)	(D, D)
$u_{R_2\|S_1}(x_1, x_2\|C, C)$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$u_{R_2\|S_1}(x_1, x_2\|C, D)$	0	1	0	0
$u_{R_2\|S_1}(x_1, x_2\|D, C)$	0	1	0	0
$u_{R_2\|S_1}(x_1, x_2\|D, D)$	0	1	0	0

Table 5 The joint praxeic selectability and rejectability for the social Prisoner's Dilemma game

$[(x_1, x_2), (x_1, x_2)]$	$\tilde{u}_{S_1 S_2}[(x_1, x_2), (x_1, x_2)]$
$[(C, C), (C, C)]$	0.3
$[(C, D), (C, D)]$	0.0
$[(D, C), (D, C)]$	0.0
$[(D, D), (D, D)]$	0.2
	$\tilde{u}_{R_1 R_2}[(x_1, x_2), (x_1, x_2)]$
$[(C, C), (C, C)]$	0.0
$[(C, D), (C, D)]$	0.2
$[(D, C), (D, C)]$	0.025
$[(D, D), (D, D)]$	0.025

The joint praxeic selectability and rejectability functions, as defined by (17) and (18) are given in Table 5, and the jointly satisficing set (with $q = 1$) is $\{(C, C), (D, D)\}$. The individual selectability and rejectability marginal utilities, as defined by (21) and (22) are displayed in Table 6, from which it can be seen that the individually satisficing sets (with $q_1 = q_2 = 1$) are $\Sigma_1^1 = \{(C, C), (D, C)\}$ and $\Sigma_1^2 = \{(C, C), (D, D)\}$. Consequently, the compromise set is $\mathcal{C} = \{(C, C)\}$.

Under the classical formulation of the Prisoner's Dilemma, the only rationally justifiable solution is mutual defection, since that formulation does not take into consideration any social relationships. From the classical point of view, mutual cooperation, although Pareto optimal, cannot be justified. The social version of the game as developed here, however, indicates that mutual cooperation is the only justified solution.

5 Reconciliation with Classical Theory

Not all problems fit naturally into the dual-utility structure of satisficing theory. One way to deal with this situation, while still retaining some of the flavor of satisficing theory, is to invoke the assumption that all consequences are rejectability neutral, and ascribe all meaningful utility to selectability.

Table 6 The individual selectability and rejectability utilities for the social Prisoner's Dilemma game

(x_1, x_2)				
$\tilde{u}_{S_1}(x_1, x_2)$	0.3	0.1	0.4	0.2
$\tilde{u}_{R_1}(x_1, x_2)$	0.25	0.25	0.25	0.25
$\tilde{u}_{S_2}(x_1, x_2)$	0.3	0.0	0.0	0.7
$\tilde{u}_{R_2}(x_1, x_2)$	0.25	0.25	0.25	0.25

Under this situation, we set the rejectability to a constant: $u_{R_j|\mathrm{pa}(R_j)}[\mathbf{a}_j|\mathrm{cp}(R_j)] = K_j = \frac{1}{|\mathcal{A}_i|}$ ($|\cdot|$ denotes cardinality) for all $\mathbf{a}_j$. We define the conditional utility of X_i as

$$u_{X_i|\mathrm{pa}(X_i)}[\mathbf{a}_i|\mathrm{cp}(X_i)] = u_{S_i|\mathrm{pa}(S_i)}[\mathbf{a}_i|\mathrm{cp}(S_i)]. \tag{27}$$

Thus, (14) becomes a function of $\mathbf{a}_i$, $i = 1, \dots, n$, only, and we may write

$$\hat{u}_X(\mathbf{a}_1, \dots, \mathbf{a}_n) = \prod_{i=1}^{n} u_{X_i|\mathrm{pa}(X_i)}[\mathbf{a}_i|\mathrm{cp}(X_i)], \tag{28}$$

and the marginals become

$$\tilde{u}_{X_i}(\mathbf{a}_i) = \sum_{\neg\mathbf{a}_i} \hat{u}_X(\mathbf{a}_1, \dots, \mathbf{a}_n). \tag{29}$$

Once all of the valuations are concentrated in a single utility, we view the decision problem from the classical perspective of optimization. The most well-known solution concept for individuals is the non-cooperative game theoretic concept of Nash equilibria [30]. Let $\mathbf{a}^* = (a_1^*, \dots, a_n^*)$. The action profile $\mathbf{a}^*$ is a *Nash equilibrium* if, were any single individual to alter its choice, its utility would decrease; i.e., if $\mathbf{a}^\dagger = (a_1^*, \dots, a_i', \dots, a_n^*)$, then, in terms of categorical utilities,

$$u_{X_i}(\mathbf{a}^*) \geq u_{X_i}(\mathbf{a}^\dagger) \tag{30}$$

for all $a_i' \in \mathcal{A}_i \setminus \{a_i^*\}$ for $i = 1, \dots, n$.

When conditional utilities are involved, we may define two notions of equilibrium. First, let us define what might be called a conditional Nash equilibrium. The action profile $\mathbf{a}^*$ is a *conditional Nash equilibrium* if

$$u_{X_i|\mathrm{pa}(X_i)}(\mathbf{a}^*|\mathbf{a}^*, \dots, \mathbf{a}^*) \geq u_{X_i|\mathrm{pa}(X_i)}(\mathbf{a}^\dagger|\mathbf{a}^\dagger, \dots, \mathbf{a}^\dagger) \tag{31}$$

for all $a_i' \in \mathcal{A}_i \setminus \{a_i^*\}$ for $i = 1, \dots, n$.

We may also compute the Nash equilibrium in terms of the marginal utility defined by (29). The action profile $\mathbf{a}^*$ is a Nash equilibrium if

$$\tilde{u}_{X_i}(\mathbf{a}^*) \geq \tilde{u}_{X_i}(\mathbf{a}^\dagger). \tag{32}$$

Example 2 (Prisoner's Dilemma, Continued) Let us revisit the Prisoner's Dilemma discussed in Example 1 under the assumption of neutral rejectability, and set $u_{X_1} = u_{S_1}$ and $u_{X_2|\mathrm{pa}(X_2)} = u_{S_2|\mathrm{pa}(X_2)}$ as defined in Tables 2 and 3, respectively. By inspection, we see that the conditional Nash equilibrium is (D, D), as with the conventional formulation. Furthermore, the marginal utilities are given in Table 7, and we see that, again the Nash equilibrium is (D, D).

Table 7 The marginal utilities for the Prisoner's Dilemma

	(C, C)	(C, D)	(D, C)	(D, D)
u_{X_1}	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{4}{10}$	$\frac{2}{10}$
u_{X_2}	$\frac{3}{10}$	0	0	$\frac{7}{10}$

The Nash equilibrium is usually considered to be an appropriate solution concept for non-cooperative games. On the other hand, with a cooperative game (i.e., one where binding agreements are possible), it may be possible to enter into negotiations and bargain for a solution. For the players to forge an agreement, however, each must achieve an acceptable degree of satisfaction. A *bargaining game* is a cooperative game in which each participant possesses a *disagreement point* that defines the benefit that is guaranteed to accrue to it if a compromise cannot be reached. The disagreement point, therefore, is an indication of the strategic strength that is conferred on the participant as it participates in negotiations: the higher the disagreement point, the greater bargaining strength of the participant.

A well-known bargaining concept that offers a clear definition of individual acceptability is the Nash bargain [29], which permits each participant to make maximal use of its strategic strength. The approach is based on four fundamental principles: (i) invariance to positive affine transformations; (ii) Pareto optimality; (iii) independence of irrelevant alternatives, and (iv) symmetry, which is the notion that no individual agent can expect that the other agents will grant it better terms than that individual itself would be willing to grant, were roles reversed.

Nash showed that these four conditions lead to a unique solution. Let d_{X_i} denote the disagreement point for X_i. The *negotiation set*, denoted $\mathcal{N}$, is the subset of action profiles such that every participant achieves at least its disagreement point. In terms of categorical utilities, the negotiation set is

$$\mathcal{N} = \left\{\mathbf{a} \in \mathcal{A}\colon u_{X_i}(\mathbf{a}) \geq d_{X_i},\ i = 1, \dots, n\right\} \tag{33}$$

and the Nash bargain is

$$\mathbf{a}_N = \arg\max_{\mathbf{a} \in \mathcal{N}} \prod_{i=1}^{n} \left[u_{X_i}(\mathbf{a}) - d_{X_i}\right]. \tag{34}$$

The intuitive interpretation of a Nash bargain is that it defines a fair compromise. It enables each player to take advantage of the strategic strength endowed by its disagreement point. The higher X_i's disagreement point, the more action profiles that are unfavorable to it are eliminated.

The structure of (34) suggests that the optimal group solution can be interpreted as a Nash bargain with unilateral utilities replaced by conditional utilities and all disagreement points set to zero. Analogously, therefore, we may define a *conditional Nash bargaining solution*. When decisions are made under certainty, the negotiation set is defined as

$$\mathcal{N} = \big\{\mathbf{a} \in \boldsymbol{\mathcal{A}}\colon u_{X_i\,|\,\mathrm{pa}\,(X_i)}(\mathbf{a})|\,\mathrm{cp}\,(X_i) \ge d_{X_i},\ i = 1, \dots, n\big\}. \tag{35}$$

The conditional Nash bargaining solution is

$$\mathbf{a}_N = \arg\max_{\mathbf{a}\in\mathcal{N}} \prod_{i=1}^{n} \big[u_{X_i\,|\,\mathrm{pa}\,(X_i)}[\mathbf{a}|\,\mathrm{cp}\,(X_i)] - d_{X_i}\big]. \tag{36}$$

Referring again to the Prisoner's Dilemma example, it is easily seen that both the conditional Nash bargain is the same as the conventional Nash bargain for the Prisoner's Dilemma; namely, the Pareto optimal solution (C, C).

6 Conditionally Decoupled Societies

6.1 The General Case

The approach developed above assumes that the conditional preferences are defined over the entire product action space. In this respect conditional preferences are generalizations of classical categorical preferences, the difference being that the preferences can be modulated by the commitments of others. Although increased complexity is associated with the introduction of conditional preferences, there are cases where this additional complexity is not justified. It can be the case that the only commitments that affect the preferences of an agent are the direct consequences to its parents. This situation motivates the notion of conditional decoupling.

Definition 5 A society is *conditionally decoupled* if the conditional preference of each agent is a function only of its own actions, given the commitments of its parents to their own actions.

Whereas, for a non-decoupled system, the utilities are functions of the entire action profile, for a decoupled system, the utilities are functions of individual actions. To develop this concept, suppose $\mathrm{cp}\,(V_i) = \{\mathbf{a}_{i_1}, \dots, \mathbf{a}_{i_{p_i}}\}$. Then the conditional utility

$$u_{V_i\,|\,\mathrm{pa}\,(V_i)}[\mathbf{a}|\,\mathrm{cp}\,(V_i)] = u_{V_i\,|\,\mathrm{pa}\,(V_i)}(\mathbf{a}|\mathbf{a}_{i_1}, \dots, \mathbf{a}_{i_{p_i}}) \tag{37}$$

becomes

$$u_{V_i\,|\,\mathrm{pa}\,(V_i)}[a_i\,|\,\mathrm{cp}\,(V_i)] = u_{V_i\,|\,\mathrm{pa}\,(V_i)}(a_i\,|a_{i_1}, \dots, a_{i_{p_i}}). \tag{38}$$

Then (14) becomes

$$\hat{u}_{\boldsymbol{SR}}(a_1, \dots a_n, a'_1, \dots, a'_n) = \prod_{i=1}^{n} u_{S_j\,|\,\mathrm{pa}\,(S_j)}[a_j\,|\,\mathrm{cp}\,(S_j)] \prod_{j=1}^{n} u_{R_j\,|\,\mathrm{pa}\,(R_j)}[a'_j\,|\,\mathrm{cp}\,(R_j)]. \tag{39}$$

The corresponding joint selectability and rejectability marginals are given by

$$u_{\boldsymbol{S}}(a_1, \dots, a_n) = \sum_{(a'_1,\dots,a'_n)} u_{\boldsymbol{SR}}(a_1, \dots, a_n, a'_1, \dots, a'_n) \tag{40}$$

and

$$u_{\boldsymbol{R}}(a'_1, \dots, a'_n) = \sum_{(a_1,\dots,a_n)} u_{\boldsymbol{SR}}(a_1, \dots, a_n, a'_1, \dots, a'_n). \tag{41}$$

We may now define a *social welfare function* as

$$W(a_1, \dots, a_n) = u_{\boldsymbol{S}}(a_1, \dots, a_n) - q_G u_{\boldsymbol{R}}(a_1, \dots, a_n), \tag{42}$$

where q_G is the joint q-value for the group. The jointly satisficing set is the set of action profiles that are jointly satisficing for the society as a whole, and is defined as

$$\mathcal{S} = \{(a_1, \dots, a_n) \in \boldsymbol{\mathcal{A}}\colon W(a_1, \dots, a_n) \ge 0\}. \tag{43}$$

This procedure, however, does not account for the possibility that the elements of $\mathcal{S}$ may not be acceptable to all (or any) of the individuals. Thus, we must also compute the individual satisficing sets. To proceed, we must first compute the selectability and rejectability marginals as

$$u_{S_i}(a_i) = \sum_{\neg a_i} u_{\boldsymbol{S}}(a_1, \dots, a_n) \tag{44}$$

and

$$u_{R_i}(a_i) = \sum_{\neg a_i} u_{\boldsymbol{R}}(a_1, \dots, a_n), \tag{45}$$

respectively. We may then define the individually satisficing sets as

$$\Sigma_i = \{a_i \in \mathcal{A}_i\colon u_{S_i}(a_i) - q_i u_{R_i}(a_i)\}. \tag{46}$$

This set includes all alternatives that are satisficing, or good enough, for X_i. The *satisficing rectangle* is the set of all action profiles such that each component is individually satisficing, and is given by

$$\mathcal{R} = \Sigma_1 \times \cdots \times \Sigma_n. \tag{47}$$

The intersection of the jointly satisficing set and the satisficing rectangle yields the *compromise set*, comprising the action profiles that are simultaneously good enough for the group and for each individual.

$$\mathcal{C} = \mathcal{S} \cap \mathcal{R}. \tag{48}$$

If $\mathcal{C} \neq \varnothing$, then we may form a *best compromise* as

$$\mathbf{a}^* = \arg\max_{\mathbf{a}\in\mathcal{C}} W(\mathbf{a}). \tag{49}$$

For many application scenarios, it will be necessary for the participants to achieve a compromise—failing to do so will not be an option. Under such circumstances, at least some players must be willing to modify their standards if $\mathcal{C} = \varnothing$; that is, there are no action profiles that are simultaneously good enough for the group and each individual at the given q_i levels. However, the satisficing approach provides a natural and systematic negotiation framework within which each individual controls the degree to which it is willing to lower its standards in an attempt to reach a compromise. By lowering its q_i-value incrementally, each X_i increases the size of its satisficing set and, hence, the size of the satisficing rectangle. By specifying the increment Δq_i that X_i is willing to reduce its standards, each participant can control the amount of compromise it is willing to offer others, if any. If enough participants are willing to lower their q_i-values sufficiently, it is easy to see that, eventually, the consensus set will be non-empty, and a best compromise can be achieved.

To protect itself from exploitation, however, each agent should possess a lower limit, q_{i_L} that specifies how much it is willing to compromise to reach an acceptable decision for the group. If all agents reach their lower limits, then the group will fail to reach a compromise. It is important to appreciate, however, that no individual is *a priori* subjugated to the will of the society in the sense that there is no possibility for that individual's preferences to receive consideration. Thus, every individual can be assured of receiving sufficient benefit, by its own definition, before agreeing to the compromise. If an individual could not enjoy at least that minimal assurance, it may not be inclined to join or remain affiliated with a society.

6.2 Social Choice

With the general multi-agent decision problem, each individual possesses its own action set. Some scenarios, however, are such that there is only one action set that applies to the group as a whole. Scenarios of this type are termed *social choice* problems. Thus, with a social choice problem, there is only one action space $\mathcal{A}$. The social welfare function (42) becomes

$$W(a) = u_S(a, \ldots, a) - q_G u_R(a, \ldots, a) \tag{50}$$

and the jointly satisficing set becomes

$$\mathcal{S} = \{a \in \mathcal{A}\colon W(a) \geq 0\}. \tag{51}$$

The individual selectability and rejectability marginals are computed according to (44) and (45), and the individually satisficing sets Σ_i, $i = 1, \ldots, n$ are given by (46). The intersection of the individually satisficing sets and the jointly satisficing set forms the *social compromise set*

$$\mathcal{C}_S = \Sigma_1 \cap \cdots \cap \Sigma_n. \tag{52}$$

If $\mathcal{S} = \varnothing$, then there is no group action that is good enough for the group and each individual. However, by reducing the q-values incrementally as discussed above, a consensus will eventually emerge. The *social consensus* set is the intersection of the social compromise set and the jointly satisficing set

$$\mathcal{G}_S = \mathcal{S} \cap \mathcal{C}_S. \tag{53}$$

The *best compromise* is the action in this set that maximizes the social welfare function; that is,

$$a^* = \arg\max_{a\in\mathcal{G}_S} W(a). \tag{54}$$

One of the most basic social structures is the human family. Thus, an important issue for the design and development of social robots is that of how to model complex family relationships. Furthermore, since family relationships are well known to virtually all cultures, they can be discussed and modeled without the need to establish detailed contextual background information. Hence, studying family structures provides a convenient and instructive vehicle with which to illustrate our theory. The following example is an anthropomorphic social analogy that demonstrates how family interdependencies can be expressed and compromise choices made using conditional utilities and satisficing decision theory.

Example 3 (The Family Walk) Suppose a family, consisting of a father X_1, a mother X_2, and a child X_3, is to take one of three possible nature walks, denoted $\{w, w', w''\}$. The father prefers long walks, the mother prefers beautiful scenery, and the child prefers an easy walk.

The first order of business in framing this scenario in the satisficing context is to settle on operational definitions for the notions of selectability and rejectability. From the point of view of each individual, the main goal is satisfaction according to its own criterion. Thus, it is reasonable to associate selectability with the degree of narrow self-interest. Accordingly, we define the three selectability utilities in Table 8.

As the operational definition of rejectability, we assume that each agent has a unit of concern for the interests of others. Let us first consider the mother. Since she has concern for the interests of her child, she will encode this information in a rejectability function that is conditioned on the selectability commitment of her child, as illustrated in Table 9.

Table 8 Individual selectability utilities

x	$u_{S_1}(x)$	$u_{S_2}(x)$	$u_{S_3}(x)$
w	0.1	0.4	0.3
w'	0.3	0.4	0.6
w''	0.6	0.2	0.1

Table 9 Mother's conditional rejectability $u_{R_2|S_1}$

x	$u_{R_2\mid S_1}(x\mid w)$	$u_{R_2\mid S_1}(x\mid w')$	$u_{R_2\mid S_1}(x\mid w'')$
w	0.0	0.4	0.5
w'	0.4	0.0	0.5
w''	0.6	0.6	0.0

Table 10 Father's conditional rejectability $u_{R_2|S_1S_2}$

	x		
	w	w'	w''
$u_{R_3\mid S_1S_2}(x\mid w,w)$	0	0	1
$u_{R_3\mid S_1S_2}(x\mid w,w')$	0	0	1
$u_{R_3\mid S_1S_2}(x\mid w,w'')$	0	1	0
	x		
	w	w'	w''
$u_{R_3\mid S_1S_2}(x\mid w',w)$	0	0	1
$u_{R_3\mid S_1S_2}(x\mid w',w')$	0	0	1
$u_{R_3\mid S_1S_2}(x\mid w',w'')$	1	0	0
	x		
	w	w'	w''
$u_{R_3\mid S_1S_2}(x\mid w'',w)$	0	1	0
$u_{R_3\mid S_1S_2}(x\mid w'',w')$	1	0	0
$u_{R_3\mid S_1S_2}(x\mid w'',w'')$	1	0	0

To interpret this table, consider the first column, which corresponds to $u_{R_2|S_1}(\cdot|w)$; that is, the child commits to selecting w. Since this walk is tied for the most preferred by the mother, she ascribes no conditional rejectability to that alternative, and places all of her conditional rejectability mass on w' and w'' in inverse proportion to her selectability. Similar arguments apply if the child commits to w' or w''.

The father's role in this decision process is first to defer to the commitments of his child, then to the commitments to his wife, and then, subject to those constraints, to reject the alternative that is least preferred in terms of his narrow self-interest. These values are provided in Table 10.

Finally, we must specify the child's rejectability. This rejectability is not conditioned, since the model does not call for the child's preferences to be influenced by the parents' preferences. Thus, the child's concern for the interests of others is neutral; that is, the child's rejectability function is uniform, as provided in (55).

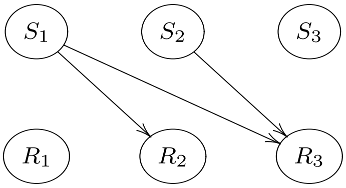


Fig. 4 A satisficing praxeic network for the family walk

$$u_{R_1}(w) = u_{R_1}(w') = u_{R_1}(w'') = \frac{1}{3}. \tag{55}$$

Figure 4 illustrates the influence flows of the satisficing praxeic network for the family walk.

Using the values provided in the above tables, we may compute the social welfare function, yielding

$$W(w) = -0.05$$
$$W(w') = 0.36667$$
$$W(w'') = -0.052;$$

hence $\mathcal{S} = \{w'\}$.

We next compute the individually satisficing sets, yielding

$$u_{S_1}(w) - q_1 u_{R_1}(w) = -0.233$$
$$u_{S_1}(w') - q_1 u_{R_1}(w') = -0.033$$
$$u_{S_1}(w'') - q_1 u_{R_1}(w'') = 0.267,$$
$$u_{S_2}(w) - q_2 u_{R_2}(w) = 0.14$$
$$u_{S_2}(w') - q_2 u_{R_2}(w') = 0.14$$
$$u_{S_2}(w'') - q_2 u_{R_2}(w'') = -0.06,$$

and

$$u_{S_3}(w) - q_3 u_{R_3}(w) = -0.12$$
$$u_{S_3}(w') - q_3 u_{R_3}(w') = 0.18$$
$$u_{S_3}(w'') - q_3 u_{R_3}(w'') = -0.32.$$

Thus, we have $\Sigma_1 = \{w''\}$, $\Sigma_2 = \{w, w'\}$, and $\Sigma_3 = \{w'\}$. Consequently, $\mathcal{C} = \Sigma_1 \cap \Sigma_2 \cap \Sigma_3 = \varnothing$, and the society has not reached a compromise that is acceptable to all participants. However, if the father reduces q_1 to 0.9, then

$$u_{S_1}(w) - q_1 u_{R_1}(w) = -0.2$$
$$u_{S_1}(w') - q_2 u_{R_1}(w') = 0.0$$
$$u_{S_1}(w'') - q_3 u_{R_1}(w'') = 0.3.$$

Hence, $\Sigma_1 = \{w', w''\}$, and a consensus exists with $\mathcal{G}_S = \{w'\}$. An important feature of this example is that the father need only reduce its standards by a small amount to achieve a consensus. In terms of the narrow self-interest offering given by Table 8, we see, after taking into consideration the social dependencies that exist among the individuals, that the consensus alternative is best for the mother and the child and second best for the father.

7 Conclusions

Multi-stakeholder decision theory, and social robotics in particular, is in need of a mathematical framework that is designed to accommodate sophisticated social behaviors such as cooperation, compromise, negotiation, and altruism. The classical framework developed by the social sciences and operations research is based on categorical preference orderings and optimization, and is not sufficiently general to characterize these social behaviors. This research represents a significant departure from classical theory by incorporating three critical notions: conditioning, coherence, and satisficing. The result is a theory of multiagent decision making that addresses the shortcomings of classical approaches and permits a reconciliation of group interests with individual interests.

In contrast to categorical utilities, which are designed to characterize self-interest, conditional utilities provide a means whereby individuals may extend their spheres of interest beyond the self. By modulating its preference structure to account for the preferences of others, an individual may account for sophisticated social relationships such as conditional altruism, and thereby give deference to others without categorically redefining its preferences.

In a homogeneous environment where decision makers are required to compromise and negotiate, it is important to ensure that no agent can be categorically subjugated. Coherence is a minimal notion of equity among the participants that can be ensured if and only if the mathematical syntax of the utilities corresponds to probability mass functions (albeit with different semantics). For societies whose inter-agent influence relationships can be represented by a directed acyclic graph, coherence ensures that the edges are conditional mass functions, resulting in a structure that is mathematically identical to a Bayesian network. This structure permits individual utilities to be aggregated to form a group utility that accounts for social relationships between individuals, thereby providing a complete model of the community.

Satisficing, as defined herein, is an approach to decision making that is as mathematically precise and formalized as is the conventional notion of optimization. The essential advantage of satisficing is that it readily extends to the multi-agent case, whereas optimization is intrinsically a single-agent concept. Furthermore, since satisficing is designed to provide a set of acceptable solutions rather than a unique best solution, it provides a natural mechanism with which to design a negotiation protocol and reach a compromise.

Acknowledgements This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under grant number W911NF-07-1-0650.

References

1. Arrow KJ (1951) Social choice and individual values. Wiley, New York. 2nd edn 1963
2. Arrow KJ (1986) Rationality of self and others. In: Hogarth RM, Reder MW (eds) Rational choice. Univ of Chicago Press, Chicago, pp 201–215
3. Balch T (2004) Hierarchic social entropy: An information theoretic measure of robot group diversity. Auton Robots 8(3):209–238
4. Balch T, Hybinette M (2000) Social potentials for scalable multirobot formations. In: Proceedings of IEEE international conference on robotics and automation, pp 73–80
5. Bergson A (1938) A reformulation of certain aspects of welfare economics. Q J Econ 52:310–334
6. Bestougeff H, Rudnianski M (1998) Games of deterrence and satisficing models applied to business process modeling. In: Proceedings of the 1998 AAAI symposium, 1998, March 23–25, Stanford California, pp. 8–14. Technical Report SS-98-05
7. Camerer C, Lowenstein G, Rabin M (eds) (2004) Advances in behavioral economics. Princeton Univ Press, Princeton
8. Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ (1999) Probabilistic networks and expert systems. Springer, New York
9. Cozman FG (2000) Credal networks. Artif Intell 120:199–233
10. de Finetti B (1937) La prévision: ses lois logiques, ses sources subjectives. Ann Inst Henri Poincaré 7:1–68. In: Kyburg HE, Jr, Smokler HE (eds) Translated as 'Forsight. Its logical laws, its subjective sources', in Studies in subjective probability, Wiley, New York, 1964, pp 93–158
11. Debreu G (1959) Theory of value. Yale Univ Press, New Haven
12. Elster J (ed) (1985) The multiple self. Cambridge Univ Press, Cambridge
13. Erlandson RF (1981) The satisficing process: A new look. IEEE Trans Syst Man Cybernet 11(11):740–752
14. Fehr E, Schmidt K (1999) A theory of fairness, competition, and cooperation. Q J Econ 114:817–868
15. Fishburn PC (1973) The theory of social choice. Princeton Univ Press, Princeton
16. Friedman M (1961) Price theory. Aldine Press, Chicago
17. Harsanyi J (1955) Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. J Polit Econ 63:315
18. Jensen FV (2001) Bayesian networks and decision graphs. Springer, New York
19. Kaufman BE (1990) A new theory of satisficing. J Behav Econ 19(1):35–51
20. Keeney RL, Raiffa H (1993) Decisions with multiple objectives. Cambridge Univ Press, Cambridge. First published by Wiley, 1976
21. Kemeny J (1955) Fair bets and inductive probabilities. J Symb Log 20(1):263–273
22. Kim Y (1999) Satisficing and optimality in 2 × 2 common interest games. Econ Theory 13(2):365–375
23. Kube CR, Zhang H (1993) Collective robotics: From social insects to robots. Adapt Behav 2(2):189–218

24. Kube CR, Zhang H (2006) Collective robotic site preparation. Adapt Behav 14:5–19
25. Mansbridge JJ (ed) (1990) Beyond self-interest. Univ of Chicago Press, Chicago
26. Margolis H (1990) Dual utilities and rational choice. In: Mansbridge JJ (ed) Beyond self-interest. Univ of Chicago Press, Chicago, pp 239–253. Chap 15
27. Matsuda T (1979) Algebraic properties of satisficing decision criterion. Inf Sci 17:221–237
28. Matsuda T (1979) Characterization of satisficing decision criterion. Inf Sci 17:131–151
29. Nash JF (1950) The bargaining problem. Econometrica 18:155–162
30. Nash JF (1950) Equilibrium points in n-person games. In: Proceedings of the national academy of sciences USA, vol 36, pp 48–49
31. Palmer FR (1971) Grammar. Harmondsworth Penguin, Harmondsworth
32. Pazgal A (1997) Satisficing leads to cooperation in mutual interest games. Int J Game Theory 26(4):439–453
33. Pearl J (1988) Probabilistic reasoning in intelligent systems. Morgan Kaufmann, San Mateo
34. Radner R (1975) Satisficing. J Math Econ 2:253–262
35. Raiffa H (1968) Decision analysis. Addison-Wesley, Reading
36. Ramsey FP (1950) Truth and probability. In: Braithwaite RB (ed) The foundations of mathematics and other logical essays. The Humanities Press, New York
37. Samuelson PA (1948) Foundations of economic analysis. Harvard Univ Press, Cambridge
38. Sen AK (1990) Rational fools: A critique of the behavorial foundations of economic theory. In: Mansbridge JJ (ed) Beyond self-interest. Univ of Chicago Press, Chicago. Chap 2
39. Sen S (ed) (1998) Satisficing models. AAAI Press, San Mateo
40. Shubik M (1982) Game theory in the social sciences. MIT Press, Cambridge
41. Shubik MS (2001) Game theory and operations research: Some musings 50 years later. Yale School of Management Working Paper No ES-14, May
42. Simon HA (1955) A behavioral model of rational choice. Q J Econ 59:99–118
43. Simon HA (1956) Rational choice and the structure of the environment. Psychol Rev 63(2):129–138
44. Simon HA (1959) Theories of decision-making in economics and behavorial science. Am Econ Rev XLIX:253–283
45. Sober E, Wilson DS (1998) Unto others: The evolution and psychology of unselfish behavior. Harvard Univ Press, Cambridge
46. Steedman I, Krause U (1985) Goethe's *Faust*, Arrow's possibility theorem and the individual decision maker. In: Elster J (ed) The multiple self. Cambridge Univ Press, Cambridge, pp 197–231
47. Stirling WC (2003) Satisficing games and decision making: With applications to engineering and computer science. Cambridge Univ Press, Cambridge
48. Stirling WC, Goodrich MA, Packard DJ (2002) Satisficing equilibria: A non-classical approach to games and decisions. Auton Agents Multi-Agent Syst 5:305–328
49. Takatsu S (1980) Decomposition of satisficing decision problems. Inf Sci 22:139–149
50. Takatsu S (1981) Latent satisficing decision criterion. Inf Sci 25:145–152
51. Tversky A, Kahenman D (1986) Rational choice and the framing of decisions. In Hogarth RM, Reder MW (eds) Rational choice. Univ of Chicago Press, Chicago, pp 67–94
52. Werger BB (1999) Cooperation without deliberation: A minimal behavior-based approach to multi-robot teams. Artif Intel 110(2):293–320
53. Wierzbicki AP (1981) A mathematical basis for satisficing decision making. In: Morse JN (ed) Organizations: Multiple agents with multiple criteria. Springer, New York, pp 465–483
54. Winter S (1971) Satisficing, selection, and the innovating remnant. Q J Econ 85:237–261
55. Zilberstein S (1998) Satisficing and bounded optimality. In: Proceedings of the 1998 AAAI symposium, 1998, March 23–25, Stanford, California, pp 91–94. Technical Report SS-98-05