

PROSPECTS FOR PREFERENCES

JON DOYLE

Department of Computer Science, North Carolina State University, USA

This article examines prospects for theories and methods of preferences, both in the specific sense of the preferences of the ideal rational agents considered in economics and decision theory and in the broader interplay between reasoning and rationality considered in philosophy, psychology, and artificial intelligence. Modern applications seek to employ preferences as means for specifying, designing, and controlling rational behaviors as well as descriptive means for understanding behaviors. We seek to understand the nature and representation of preferences by examining the roles, origins, meaning, structure, evolution, and application of preferences.

Key words: decision theory, preference, utility, rationality, limited rationality, reasoning, learning, preference representation, preference structure, preference change.

1. INTRODUCTION

The notion of preference, formerly of concern primarily to philosophers, economists, and psychologists, has in recent years drawn new attention from students of artificial intelligence. The new studies follow earlier work in using formalized preferences to understand and characterize agents of limited rationality, but go beyond the earlier studies by using formalized preferences as means for designing, instructing, and controlling such agents as well. These new investigations place additional theoretical and practical burdens on the preference formalizations.

In the following, we examine some prospects for preferences in light of problems generated by the new burdens of agent design and control. We claim that proper understanding of preferences requires acknowledging that the origin and purpose of preferences matter to their structure, that preferences play many roles in reasoning, that different roles call for different representations, and that preference changes pervade deliberation and action. If true, these claims leave us with many more questions than answers regarding preferences. We organize the discussion around the fundamental questions of the nature, roles, structure, representation, and evolution of preference. We do not attempt a complete survey of the field or the literature, but instead present a selection of topics that reflects a personal view of the issues. The present article builds on the earlier introduction (Doyle and Thomason 1999).

1.1. Preference Order and Utility

Economists formalized the notion of preference as a way of characterizing the essence of philosophical concepts of utility. The resulting formalization of preference has since become a foundational concept of the mathematical theory of decision making (Luce and Raiffa 1957; Jeffrey 1983; Gärdenfors and Sahlin 1988). In barest essentials, economists say preferences of an agent, whether those of an individual person or of a group of persons, consist of a complete preordering of a set Ω of individuals called *outcomes* or *alternatives* by a reflexive and transitive relation \preceq of *weak preference* over the elements of Ω . The reflexive and irreflexive components of weak preference in turn constitute relations \sim of *indifference* and $<$ of *strict preference*.

For many purposes, it is useful to study utility functions $U : \Omega \rightarrow \mathbb{R}$ that constitute numeric representations of preference relations. We say that U represents \preceq just in case we have $U(x) \leq U(y)$ iff $x \preceq y$ for every $x, y \in \Omega$. When only the order matters in representing preferences, we call the utility function *ordinal*. Clearly if U represents \preceq , then so also does

every monotone increasing transformation $U \mapsto m(U)$ of U , where we say that m increases monotonically just in case $m(x) \geq m(y)$ whenever $x \geq y$.

The notion of *cardinal* utility arises when magnitude matters as well as order, as it does when considering expected utility of actions. Expected utility involves preferences over *prospects*, where prospects consist of probability distributions over possible states of affairs that represent uncertainty about the outcomes of actions. Expected utility weights the utility assigned to each outcome by the probability of that outcome. Basing action on this linear combination of probability and utility restricts the flexibility of utility representations to positive affine transformations. Specifically, in cardinal utility measures, the units and origin of utility do not matter, and every positive affine transformation $U \mapsto aU + b$ of U (positive means $a > 0$) of a utility function U represents the same notion of cardinal utility.

1.2. Reasons for Re-examination

Although the standard economic conception of preference remains adequate for traditional purposes, the economic notion has become increasingly unsuitable as a basis for new applications due to the focus of economists on “nice” preferences satisfying the following properties:

- *Coherent* preferences: The agent’s preferences fit together transitively, in that the agent prefers a to c whenever it prefers a to b and b to c .
- *Consistent* preferences: The agent does not prefer a to b at the same time as it prefers b to a .
- *Complete* preferences: For each a and b , the agent either prefers a to b , prefers b to a , or holds them equally desirable.
- *Comparatively constant* preferences: The agent’s preferences do not change from one instant to the next, or else change very slowly with respect to decisions and actions.
- *Concrete* preferences: The agent’s preferences (as opposed to multiattribute representations) relate concrete alternatives, not abstract qualities or properties of alternatives.
- *Conveniently cleaved* preferences: The agent’s preferences divide along convenient conceptual lines, admitting additive representations over independent dimensions.
- *Completely comparative* preferences: Preferences serve only to indicate comparisons between alternatives in making choices.

Although these properties provide a good foundation for strong and elegant economic theories and for the analysis of many practical problems, one need not let economists monopolize the notion of preference for their own purposes. Instead, today one sees the need for a variety of theories of preference aimed at serving different purposes.

Consider, for example, artificial intelligence work on automated decision formulation and decision-theoretic planning. At times one looks for good means of specifying information about preferences by constraining the preferences held or adopted by an agent. One might well satisfy this need by using a logical language to state *qualitative* characteristics of or constraints on the agent’s preferences. In contrast to concrete economic preferences, such generic or qualitative preference statements would provide the automated reasoner with broadly applicable relations among many alternatives (“prefer air transport to rail or automobile”) rather than mere relations among specific alternatives (“prefer today’s noon United Airlines flight from LaGuardia to today’s 7 AM Amtrak train from Penn Station”). At other times one looks for good means for using the preferences so specified to make decisions efficiently. One might want a numeric utility function to satisfy the latter need. Automated decision formulation and decision-theoretic planning thus can benefit from extending traditional concepts with logical languages for specifying preferences, from automated means for

inferring properties of preference specifications, and from automated means for constructing utility functions that satisfy preference specifications (see, for example, McGeachie 2002).

Consider, for another example, work on user modeling in human–computer interaction. Although one can think of preference modeling as adjusting a single utility model in light of evidence about the subject’s preferences (see Chajewska and Koller 2000; Chajewska, Koller, and Parr 2000), this approach can pretend to represent more information about the subject than the modeler actually has by not clearly distinguishing what the modeler knows from what the modeler assumes. For instance, the modeler might know the subject views two alternatives as equally good, or might have no information about the subject’s views on the two and equate their desirability to obtain the simplest model. In the one case, new information that calls for splitting the two alternatives contradicts earlier presentations by the subject; in the other case, the same information does not indicate any inconsistency on the part of the subject, but instead represents a positive increase in the knowledge of the modeler.

Work on automation of deliberation and action in novel circumstances reveals many reasons for overturning the ideal economic properties in favor of more realistic conceptions of preference. In fact, automated deliberation and action provide a concrete context for expanding a search long conducted by economists themselves, especially in the area of multi-criteria decision making (Keeney and Raiffa 1976; Arrow and Raynaud 1986), that aims to bring the perfection of standard ideal rationality closer to more realistic ideals by relaxing one or more of the properties characterizing preferences, beliefs, and rational decisions. This search is not easy. Few like giving up beauty and simplicity for complexity, and one often feels some disorientation abandoning familiar regions for less familiar ones. Nevertheless, although the economic model of preferences has proven very fruitful, and continues to serve well in many studies, the time has come to venture beyond this familiar venue and seek out new formalisms better suited to the broader conception of preference.

1.3. Illustration

Negotiation offers a setting for examining preferences that illustrates a number of the issues discussed in the following, including the use of preference expressions to provide succinct instructions for controlling behavior, and the ease with which preferences can change over time. Our discussion here draws on work with Robert Laddaga and Vera Ketelboeter of MIT concerning the practices of the aircraft maintenance crews of a U.S. Marine Corps Air Group.

When a Marine aircraft maintenance crew needs a replacement part not on hand in the maintenance bay, its maintenance chief first seeks to obtain the part from a standard parts inventory system operated by the Marines. If that standard inventory cannot supply the part quickly enough, the chief turns to negotiation with other potential suppliers. The chief first seeks to negotiate for the part with the chiefs of the maintenance crews of other squadrons. If those negotiations fail to yield the needed part, the chief might turn to commercial suppliers.

Although some simple negotiation processes, such as those conducted by the standard proxy bidding mechanisms on Ebay, might involve only fixed preferences, the negotiations conducted by the Marine maintenance crews consist of multiple stages, with very different preferences guiding actions in different stages of the negotiation. In the negotiations with sister squadrons, the chief employs preferences reflecting the collaborative relation between the squadrons. Although each squadron has responsibilities separate from those of other squadrons, each squadron in the group serves the same commander, who judges each squadron primarily in terms of service to the shared group aims and commander’s mission. Negotiations with sister squadrons thus concentrates on bartering parts and trading favors for future redemption. All the chiefs involved know the shared mission, and one will give up

parts necessary to his own tasks if those parts are necessary to permit some other chief to enable some more important part of the shared mission. The bartering thus does not seek to extract the last ounce of benefit from the trade. In contrast, negotiations with outside suppliers go much more along the lines of looking for the best price and delivery time. The chief shares no aims with the outside suppliers, and has no special interest in furthering their profitability. In these negotiations, the chief might look to get the lowest price possible.

To automate such negotiation processes, one looks to formalize both negotiating preferences and changes of negotiating preferences appropriate to the different stages or contexts of negotiation. We view the parts negotiations just discussed as proceeding through a sequence of “positions” characterized by sets of acceptability constraints on and preferences over potential agreements. This type of *constraint-based* negotiation procedure involves both “objective” and “positional” preferences. Briefly, objective preferences control whether, when, and how to bid in parts negotiations, as well as when to accept bids received, while positional preferences control when and how to change the objective preferences.

In this view of constraint-based negotiation, negotiating positions consist of a set of intended conditions or constraints on acceptable deals together with a set of preferences over possible deals.

Objective search consists of using the preferences over possible deals ranked within in the current position—the objective preferences—to search for the best deal within constraints of the current position by making bids and accepting or rejecting proposals from other participants.

Positional search, in turn, consists of search for a position offering the best or an acceptable deal. Positional search varies constraints on and preferences over deals when needed deals seem unavailable or infeasible within the current position. Variations can include weakening or strengthening either constraints or preferences contained in the position.

Just as objective preferences over deals guide objective search, positional preferences over negotiation positions guide positional search. Although these preferences can relate positions by comparing global properties (e.g., consistency or completeness) of the positions, positional preferences typically compare the individual constraints and preferences contained in the positions. For example, one often interprets “soft” constraints in terms of preferences expressing relative desirability of specific constraints. One can consider a variety of forms of such preferences among constraints, including

- Simple preferences $c_1 < c_2$ that indicate that one constraint (c_1) is more desirable to maintain than another (c_2);
- Compound preferences $\{c_1, c_2\} < \{c_3\}$ that indicate that maintaining one set of constraints ($\{c_1, c_2\}$) is more desirable than maintaining another set ($\{c_3\}$);
- Conditional preferences $c_1 \dashv\vdash c_2 < c_3$ that restrict the stated preference ($c_2 < c_3$) to circumstances in which specific conditions (c_1) obtain; and
- Nonmonotonic preferences $c_1 \parallel c_2 \dashv\vdash c_3 < c_4$ in which the circumstantial restrictions ($c_1 \parallel c_2$, read “ c_1 without c_2 ”) refer to absence of conditions (c_2) as well as to their presence (c_1).

Simple positional preferences of these sorts thus correspond to the entrenchment orderings identified by Gärdenfors and Makinson in the context of belief revision (Gärdenfors and Makinson 1988), or the more general rational revision preferences over beliefs, preferences, and other mental attitudes examined by Doyle (1990, 1991). The full range of positional preferences correspond to these more general preferences, as they include preferences over preferences of forms corresponding to the preferences over constraints, such as the simple preference $(c_1 < c_2) < (c_3 < c_4)$. Gordon’s (1995) example from the law illustrates similar preferences nicely. The principle of *Lex Posterior* states that in deciding judicial cases

later laws should be preferred to earlier laws as the basis for decision. The principle of *Lex Superior* states that laws from higher authorities should be preferred to those of laws from lower authorities. These preferences can sometimes conflict; in such cases, one prefers to follow *Lex Superior* to *Lex Posterior*. Positional preferences often relate to or reflect objective preferences, but need not always do so.

With this view of constraint-based negotiation, negotiation strategies consist of staged or randomized changes of negotiation position. Although one can construct negotiation strategies through a rational planning process, negotiation plans can require revision just as frequently as plans for other activities. One thus can expect realistic methods of constraint-based negotiation to involve midstream rational adjustments to or incremental construction or adaptation of an evolving strategy.

The negotiation processes just sketched illustrate two important points. First, preferences change, so one should not base a practical approach on a theory that presupposes fixed preferences. Of course, one can sometimes prescribe patterns of change in advance by identifying specific dependencies on particular aspects of the context in which decisions will be made, but such foreknowledge is not always available. Second, some preferences concern mental entities, such as intentions and even other preferences, so one should not base a practical approach on representations that hide preferences over propositions deep within an inscrutable numerical appearance. In the following, we examine these points in more detail, along with other issues underlying the theory and practice of preferences in reasoning and action.

2. NATURE

What are preferences?

Economists generally believe this question has been answered by the economic theory of preference, which identifies preferences as relations between alternatives exhibiting the structure of a complete preorder. Economists regard the problem of choosing among alternatives as determining the essential nature of preference. Focusing attention on this narrow problem then lets them determine the nature of economic preference in exclusively comparative terms, that is, as a pure ordering relation.

The economic conception of preference has much to recommend it. The exclusively comparative nature it presumes yields a very simple formal concept. Moreover, this conception lets one read preferences into choices. Using methods developed by Von Neumann, Morgenstern, and Savage, the economist reconstructs the preferences of a rational individual by examining sets of choices made by the agent. This involves interpreting actual or hypothetical choices put to the agent, choices that serve to pin down the ways the agent would act in all circumstances. The interpreter typically must obtain infinite numbers of such hypothetical choices, but one can reduce this burden if one can shrink the scope of action to finite numbers of alternatives and outcomes.

Despite these positive characteristics of the economic conception of preferences, the economist's answer does not really answer the question, but instead begs it by presuming the existence of rational individuals of a character compatible with the economist's formalization of preferences. In a real sense, the economist defines the notion of economic agent in terms of the notion of economic preference, rather than the other way around.

But are preferences only economic preferences? Put another way, are the preferences we attribute to people, organizations, and even pet cats really preferences in the economic sense? Because preferences are defined independently of rational individuals, answering this new question means looking at the sorts of agents that might have preferences to see whether they do or can have preferences of the sort identified by economists.

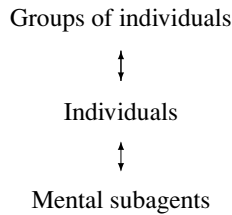


FIGURE 1. Three levels of agent types. The arrows indicate that preferences of agents at each level can shape preferences at other levels.

In fact, even economists do not claim that rational individuals exist, at least when employing the economist's characterization of rationality, because people sometimes fall short of the requirements of the economic ideal. Economists therefore only identify rational individuals as an ideal type of agent approximated to some degree by people. The question is whether the economist's characterization of preference survives these approximations, or whether characterizing human preferences and reason in useful ways requires one to depart systematically from economic rationality. That is, do the ways that actual agents of interest diverge from the economic conception still allow us to say that economic preferences actually exist, or is the economic conception so different from any realistic agent as to render the conception useless for practical specification and analysis?

Economists consider preferences exhibited by individual persons or by groups of individuals that act as if they were persons. Artificial intelligence broadens the range of possible preferrers to artificial automated agents, including both automated rational decision makers of the sort identified by economists (see, for example, Russell and Norvig 2002) and individuals with minds composed of societies of mental subagents or other substructure (see, for example, Minsky 1986). Persons and groups and mental subagents form three levels of types of agents, depicted in Figure 1, to which one might ascribe preferences. In what sense can agents at these different levels have preferences in the economic sense?

For agents consisting of groups of individuals, Arrow's impossibility theorem (Arrow 1963) casts strong doubt on the existence of rational group preferences except in special cases. Arrow identifies a few general and natural relations one expects to hold between individual and group preferences.

1. *Collective rationality*. Each arrangement of individual preference orders determines a unique group preference order.
2. *Pareto principle (unanimity)*. The group order agrees with uncontested strict preferences, so that any disagreement between the group order and some individual order implies a disagreement between that individual and some other individual.
3. *Independence of irrelevant alternatives*. The relation of two alternatives according to the group order depends only on how the individuals rank those two alternatives, so that the group order restricted to a subset of candidates yields the same result as an aggregation of the individual orders restricted to that subset.
4. *Nondictatorship*. There is no individual (the "dictator") whose preferences automatically determine the group preferences, independent of how other individuals order the alternatives.

As these requirements indicate, Arrow treats only relations concerning the influence of individual preferences on group preferences, not relations concerning the influence of group preferences on individual preferences. He then shows that there is no way of deriving rational

group preferences from the individual ones that satisfy the relations he identifies. The main paths around this result involve restricting the character of the individuals or permitting one individual to dictate the entire group preference order. Neither of these approaches seem obviously preferable to questioning the whether preferences as exhibited by groups really have the character required by economists.

Even though group preferences do not seem to have the character identified by economists, perhaps the humans making up such groups do have preferences in the economic sense. After all, economists developed their axioms for preferences to idealize human thought and action. Unfortunately, appreciation of the economist's definition provides little reason to assume people have preferences in the economist's sense. Almost every reflective person knows that the grounds for personal decisions are not complete. One continually faces new, heretofore unimagined, decisions. The grounds for personal decisions prove fluid, not fixed, even in people who are not fickle, especially when different concrete situations highlight conflicts between fragmented and seemingly irreconcilable values (see Van Fraassen 1973; Nagel 1979; Harty 1994). In complex automated agents, one finds that the same grounds for decision can yield distributions of different decisions over time, either because making a decision requires inference and construction (Doyle 1983a, 1989a), or because the agent's actions represent an unstable decision by a group of subagents subject to the constraints of Arrow's theorem (Arrow 1963; Doyle 1985; Doyle and Wellman 1991). These characteristics of the decisions of actual agents make a hash of the standard economic analysis of preference in which one assumes an agent can be forced to make an infinite number of decisions without change. One may assume one can do that to an agent, but the assumption is great.

Even if one is tempted to conclude people do not have preferences in the economic sense, one expects that some agents might have preferences. One can certainly lay out the design for artificial agents simple enough to have identifiable preferences. For example, one easily imagines that some of Minsky's mental subagencies (Minsky 1986), operating to seek to achieve one specific goal, can be characterized perfectly adequately as having preferences in the economic sense. However, such a narrowly focused mental subagent is not just simple, but is *really* simple, of intrinsic complexity comparable to the thermostat to which McCarthy (1979) ascribes belief. Do such seemingly tiny mental subagencies provide the only realistic setting in which economic preferences could exist?

Put another way, can suitably simple persons also have preferences? The worry here is that properly calling an agent a person might entail a psychological structure too rich to ensure the existence of preferences in the economic sense. In particular, Frankfurt's (1971) characterization of persons with free will involves a nontrivial hierarchy of reflective preferences and potential conflicts among these. In this picture, preferences at the individual level can modify or qualify those at the subagent level. Deliberation conducted with such reflective preferences, as in (Doyle 1980), generates in the conduct of nonmonotonic reasoning difficulties resembling those arising in the case of rational group decision making (Doyle and Wellman 1991).

The possibility that artificial agents might be endowed with economic preferences does not mean one would always want artificial agents to have preferences in the economic sense, especially if one wants such agents to resemble people. The Red Queen of Alice's wonderland prescribed decapitation for those pointing out her inconsistencies. Real people might not wish the same for spouses or teenagers who object to apparent inconsistencies, but they might at least sometimes wish for toleration, acceptance, and even sharing of inconsistencies. In similar way, many might want artificial agents with which they work to exhibit at least some of the inconsistency and incompleteness afflicting human persons.

The economic conception of preference endows preferences with a nature solely derived from their role in comparing alternatives. Assuming that preferences have this nature

provides important theoretical benefits, but at the same time means that the economic conception provides little guidance in understanding preferences in other terms, and especially in understanding how to approach violations of economic assumptions. To look for deeper understanding, one must look to other roles of preferences in reasoning and action, and to the sources or origins of preferences.

3. ROLES

What do preferences do?

To the economist, the primary role played by preference is in choosing rationally among alternatives. Economists define the notion of preference in terms of rational choice (and vice versa), so that one must assume the rationality of the agent under interpretation to interpret actions as giving information about the agent's preferences. If the agent is not rational, this interpretation process either constructs preferences for some rational agent whose actions could match the observations, or determines that no such rational interpretation exists, that is, determines that either the observed agent is not rational or that it has changed its preferences during the period of observation.

Even though the economist's notion of preference is inseparable from the economist's notion of rationality, one can look for preferences of other characters by examining other roles that preferences can play, such as providing convenient means for exercising control over reasoning; constituting impermanent elements or aspects of mental states; serving as constructors of mental states; forming objects of reasoning; and constituting the content of human thought and action. Just as economists use the role of preference in rational decision making to interpret decisions in terms of preferences, one can use the role of preference in these other activities to interpret such activities in terms of preference, looking at the effects or influences of preferences in making reasoning faster and introducing new decisions in addition to the traditional effect of selecting between alternatives. The form of such roles and the methods of interpretation are more varied and complex than in the economist's case, but the idea is the same.

3.1. Exercising Control

First of all, economists recognize the normative use of preferences as well as the purely descriptive use appearing in the standard elicitation methods. In this role, stipulated preferences provide part of the means by which one can control an artificial (or human) agent. This role raises issues beyond the purely comparative conception favored by economists, for in it preferences act as tools, and people demand a degree of convenience and utility in their tools. In the case of preferences, this means providing methods for specifying preferences that go beyond concrete preorders and multiattribute utility functions. Following experience in specifying taxonomic and epistemic information in artificial intelligence, one looks for graphical, qualitative, and logical specification languages that permit provision of information for controlling reasoning and action in natural terms (de Kleer et al. 1977; McDermott 1978; Doyle 1980; Doyle and McGeachie 2003). Natural specifications of control information complement or supplant traditional multiattribute specifications with methods for stating goals and stating generic orderings. Convenience also calls for methods that permit the designer to state simple adaptations simply. For example, one easily can state changes in goals and generic orderings ("Ask not what your country can do for you—ask what you can do for your country.") that correspond to no simple change in standard utility functions.

3.2. Elements of Mental States

Preferences play a second role as elements of mental states. Preferences, like other mental attitudes, can form ephemeral or changeable elements that come and go in the course of reasoning and action. They can also form constitutive elements of mental states that provide an unchanging basis for reasoning and action.

This role figures prominently in so-called BDI agents in which the mental attitudes of beliefs (B), desires (D), and intentions (I) mediate the agent's reasoning and action (Rao and Georgeff 1991). One naturally views preferences as comparative or relative desires reflecting the comparative desirability of one thing over another. The exact nature of preferences in a BDI agent varies with the agent architecture and the processes of deliberation, volition, memory, and perception used by the agent to reach decisions or actions from mental attitudes (Doyle 1988, 1989b). Different agent architectures can employ different deliberative processes, and thus involve different conceptions of preference. Here we refer only to differences in the roles played in similar processes, not to differences in representation of preferences, though architectures normally employ representations tailored to the use made of them.

For example, the early BDI architecture of Doyle (1980) used at least two different forms of preferences in different subprocesses, with one fairly explicit form acting as the expression of comparisons among alternatives in the standard decision-theoretic way, and with another form of preference implicit in nonmonotonic reasons or justifications acting as guidance for making assumptions and revisions. This same architecture carried within it provision for an array of different deliberative and volitional processes. Each of these would in turn identify preferences differently, as do the many other architectures set forth in the literature.

Defining preferences with respect to their role in psychological systems or processes merely follows the pattern of the standard economic conception, albeit with a broadening of perspective to include other possible psychologies or organizations for minds (see Doyle 1982) than that of the rational economic agent, and other mental processes than making choices. Rather than deriving the nature of preferences only from the role played in comparing alternatives, the broader conception supports formal characterizations in terms of Ramsey sentences (Ramsey 1931)

$$\exists x \forall \vec{y} R_{[\text{psychology, process}]}(x, \vec{y}),$$

read “there exists something (x , called preference) that plays the role or holds the relation R to other elements (\vec{y}) in the psychology and process”, that pick out preference by reference to the overall role played in the mental organization and process. For the economist, the psychology is always that of the rational agent; the process is irrelevant; and the relation is always that of selecting alternatives of maximal expected utility.

3.3. Constructors of Mental States

If the second role highlights the appearance of preference as changing elements of mental states, the third role emphasizes a related role as constructors of mental states (Doyle 1989a). This role to some extent combines the first two by using preferences to form and inform reasoning policies that represent rational decisions about what to think of the sorts involved in Pascal's wager (Pascal 1962) and James' will to believe (James 1897).

This role finds expression in preference-based theories of nonmonotonic reasoning (Doyle 1983b, 1985, 1994; Shoham 1987). In these theories, individual reasons serve as distinct sources of preference over mental states. For example, one interprets a nonmonotonic reason or default rule $A \setminus B \Vdash C$ (read A without B gives C) as expressing a preference conditional on the antecedents (A) for states containing the conclusions (C) over states

indicating defeat (B) and for defeat states over states reflecting agnosticism (neither B nor C) (Doyle and Wellman 1991). Standard definitions of nonmonotonic consequences of sets of such rules then turn out to produce Pareto-optimal group choices with respect to the individual rules, that is, sets of conclusions that cannot satisfy one rule more without satisfying another rule less. Pareto-optimality of the sets of conclusions encourages an interpretation of such conclusion sets as economic equilibria (Doyle 1985). These choices based on sets of reason preferences resemble the group decision problem for sets of people, and a variant of Arrow's theorem rules out any easy notion of universal default logic. Delgrande et al. (2004) survey the ways in which recent nonmonotonic formalisms incorporate preference information.

Such preferences over attitudinal states also appear in the positional preferences in the negotiation illustration, where they work to construct mental states at different times rather than at a single time.

3.4. Objects of Reasoning

Preferences play a third role as conclusions of reasoning. Reasoning toward preferences forms a central element of reasoned deliberation (Doyle 1980). In reasoned deliberation the agent argues with itself about which alternatives are preferable to others. For example, the agent might arrive at some preference by means of reasoning similar to the following:

- P is better than Q because *[some reason]*.
- On the other hand, Q is better than P because *[another reason]*.
- Yet P is still better than Q because *[a third reason]*.
- I can't think of any further arguments.
- Therefore P is better than Q.

Reasoning toward preferences thus can resemble reasoning toward belief.

Although this role bears close connections to the second role as elements of mental states, it seems worth distinguishing as a distinct role. Although reasoning to preferences as conclusions is easiest to consider when the conclusions in question are changeable elements of mental states, the mere fact of changes in the composition of mental states does not entail that these changes represent reasoned conclusions. Changes to the mental elements constituting a state instead might occur as effects of actions or perception. Moreover, reasoning to preferences does not imply that preferences represent changeable elements of mental states. One might reason to some preference as a conclusion even though one's preferences do not take the form of distinct elements of one's mental state, and even though one lacks a direct means for effectualizing the preference. For instance, one might conclude "I should like broccoli" even if the only means toward realizing this conclusion is to force oneself eat broccoli frequently enough that the resulting familiarity induces a favorable attitude.

The role of preferences as objects of reasoning also bears close connections to the third role as constructors of mental states. Of course, one might have preferences construct mental states that do not themselves include preferences as distinct objects, and state-constructive preferences might not refer to specific preferences as opposed to constellations of attitudes. On the other hand, some possible organizations for minds blur the distinction. For example, explicit preference statements can stand as the conclusions of arguments constructed through reasoned deliberation like that sketched at the beginning of this subsection. Indeed, even the reasons underlying those arguments admit treatment as conclusions and subjects of other reasons (see Doyle 1980, 1983b).

Reasoning toward preference plays a critical role in dealing with novel circumstances. (“We seem to be in an uncharted system, Captain. Should we head to that planet or that moon?”) In such circumstances, one lacks preferences directly applicable to the available alternatives and must use reasoning or other means to derive or invent relevant preferences.

Preference conclusions also figure in learned preferential generalizations, in which one abstracts from concrete experiences to more general statements of preference. (In the immortal example of Theodore Geisel, “Say! I like green eggs and ham! I do! I like them, Sam-I-am!”) Learned preferential generalizations resemble the standard von Neumann–Morgenstern–Savage economic methods of inferring preferences from choices in that they start with experience and end with preferences. They differ in that the economic methods work only with specific choices involving choices between clearly formulated alternatives. More general learning methods might yield inferred preferences even when the experience involves choices or behavior of a much wider character.

3.5. Content of Thought and Action

The final role for preferences we consider is as the content of human thought and action, and specifically, as objects of communication, as things people tell each other or tell machines. This role for preference has much in common with that of the role as objects of reasoning, but with defining or constraining characteristics derived from the communication medium and the communicants. In telling each other what they like and dislike, humans clearly suffer great restrictions on the length and complexity of communications in comparison with artificial agents transmitting large representations across high-speed networks. In part due to these limitations, people communicate preferences to each other using both verbal and nonverbal means.

Nonverbal communication of preference includes bodily motion and action patterns. For example, one might signal preferences and likes by hovering near something, by widening one’s eyes in considering it, or by jumping to attention when it appears. One might signal contrary preferences and dislikes by backing away from something, wrinkling eyes when considering it, and by ignoring requests concerning it. Indeed, one can view any explicit choice, verbal or not, as behavior indicating preference, as is done in economics to interpret standard gambles as preference information.

Verbal communication of preference includes both direct and indirect expression. Direct expression of preference information includes words such as “prefer,” “like,” “desire,” and “want”. Indirect expression includes statement of plans, arguments, and choice of connotation or metaphor. For example, answering a query about which road to take by saying “I plan to take the highway” might express a preference, as a linguistic statement of physical choice. As another example, continued argument for some alternative often indicates some preference for the alternative. Finally, on hearing or reading “Shall I compare thee to a summer’s day?,” no one expects the poet to continue with derogatory remarks about the subject.

Although linguists surely know more about direct expressions of preference information than is commonly known among researchers on preference, even they are unlikely to have cataloged the range of possible expressions people can use in conveying preference information implicitly, say in describing plans and activities. Can one catalog or classify all preference expressions? An enumeration might be impossible, as there might be as many ways of expressing a preference indirectly as there are objects that might enter into such an indirect expression. One might then look for more piecemeal characterizations. Can one recognize their presence in some more complex expression? Can one pick them out within more complex expressions?

4. ORIGINS

Where do preferences come from?

Preferences can derive from many sources or origins. One expects that preferences arising from different origins will exhibit different characters, so a thorough understanding of the nature of preference entails understanding special structures arising from these different sources.

Reflection on psychology offers numerous possibilities for mental entities that can produce or influence preferences, including desires, drives, motives, sentiments, emotions, will, intentions, plans, habits, and instincts. Each of these might generate or shape preferences in different ways. The early work of Shand (1920) cataloged a variety of mental affects and the way they influence action, including some that relate to preference. James (1890) provides further discussion of the relation between various psychological entities and preferences. Later psychologists have claimed human nature involves six (or some small number) of fundamental drives from which all motivation stems (Bolles (1975) surveys the area). However, even these theories do not suppose the influence of such drives on motivation is direct in all cases. Instead, experience and education build on or modify these initial or underlying motivations, introducing a hierarchy of new motivations that can shape action independent of (and sometimes in conflict with) the “primal” drives that first gave them life. In this way, an original drive for food might come to produce a derived drive for sweets that, when indulged for long enough a period, produces more specialized drives for ice cream, or even ice cream topped with fresh blueberries. Indeed, traditional utilitarian philosophy mirrored this sort of hierarchical approach by positing an ever expanding hierarchy of pleasures or happiness learned over time.

One also can view the several psychological sources of preferences as complemented by moral, pecuniary, and biological sources as well. Typical moral preferences turn the agent toward good and away from evil. Typical pecuniary preferences turn the agent toward material wealth and away from material poverty. Typical biological preferences include physical hungers, but can also extend to abnormal effects of metabolism on thought. One might regard these sources as mere instances of psychological desires, but the noninterconvertibility of these origins is proverbial.

Does this variety of origins really matter? Philosophers, for instance, tend to use the term “desire” to cover a variety of sentiments and attitudes, including likes and dislikes, loves and hates, motives and drives. One can sympathize with a philosophical urge to abstract all these origins to a general notion of desire, but doing so risks losing important insight into the nature and structure of preference. Most adults, for example, have intimate acquaintance with the possibility (some would even say inevitability) of loving and hating the same thing. This commonplace sounds odd when translated into the language of desire. In translation, the love and hate then sound like desiring and not desiring something simultaneously, which a logician uncommitted to the human truth might simply reject as nonsense instead of trying to understand it as reality.

Thus the special structure of various sources for preference raises the possibility of special structure obtaining in the preferences resulting from each source. It seems foolhardy to ignore such additional information about preference if it exists. Studies of preference, however have only looked at a few of the potential origins of preference in any detail. These special cases include the standard economic notions of risk aversion and wealth effects and the psychological notions of framing and baselines (Kahneman, Slovic, and Tversky 1982; Machina 1987). Artificial intelligence contributions include Haddawy and Hank’s (1998) study of preference arising from deadlines and time pressure and Wellman and Doyle’s (1991) identification of the preferences generated by goals (or vice versa).

Formal connections between goals or other psychological phenomena and preferences need not presume that one generates the other. As the work of Wellman and Doyle (1991) indicates, one can use formal definitions and a given set of preferences to identify the goals of the agent holding the preferences, or use the definitions and a set of goals to identify preferences held by the agent. One can imagine identifying similar relations of preference with some of the more specific sentiments and attitudes.

Unfortunately, the study of motivation has tended away from looking at specific origins in favor of seeking formalisms of increasing generality. The economist's notion of preference *simpliciter* constitutes one terminus of this path. Other investigators end up with similarly general pseudo-physical treatments of utility representations as potential fields akin to gravitational or electromagnetic potentials. These abstract frameworks undoubtedly bear further study, but one need also remember that more specialized paths have not attracted the attention they deserve.

5. REPRESENTATION AND MEANING

What do people mean when they express preferences?

Most methods for engineering the preferences employed by agents or in decision models rely on expressing the preferences or information about the preferences in a formal or informal language. The experience of artificial intelligence suggests that we need not restrict such languages to the most elemental logical primitives, but should exploit the added power, succinctness, and convenience of more complex vocabularies.

In choosing expressive languages for specifying preferences, one can look to the expressions and methods used by people to communicate preference information, as discussed earlier in Section 3.5. One need not look deeply at such expressions to notice that intended meanings vary even when people use the same words. Thus employment of human preference expressions introduces a variety of preference types on a par with the variety resulting from different origins for preferences.

5.1. Semantical Varieties

Formal semantics for preference logics have already been developed that encompass a variety of meanings for linguistic preference expressions. The formal semantics capture differences in detail quite invisible in reading dictionary definitions of "preference", but quite significant when considering the inferential power each meaning variant provides in a preference logic.

The simplest formal semantics studied to date captures the straightforward language of order relations on individuals and simple lifting to sets of individuals. This semantics captures much of common usage for individuals, but relatively little common usage when applied to classes of individuals. Formally, the application to classes trivializes all tradeoffs to indifference (see, for example, Wellman and Doyle 1991).

The next formal semantics captures the multiattribute utility functions common in everyday applications of decision theory (Keeney and Raiffa 1976). This semantics presupposes a set of attribute sets A_i indexed by a set \mathcal{I} ; a corresponding assignment of values $\vec{x} = (x_1, \dots, x_n)$ in $A = \prod_{i \in \mathcal{I}} A_i$ to individual outcomes; and a utility function $U(\vec{x})$ that identifies the utility of an outcome x in terms of the attribute values assigned to x . This provides for expression of tradeoffs among different qualities of alternatives, especially when one can construct the overall utility function by composing simpler "subutility" functions defined over subsets of the attributes, which in turn imply the independence of some sets of

attributes with respect to the contributions of these sets to utility (Wellman and Doyle 1992). The framework of multiattribute utility functions does not in itself provide a convenient language for characterizing these tradeoffs. It also applies only to individuals, not to categories of individuals.

Ceteris paribus preference semantics goes beyond plain ordering and multiattribute specifications by combining elements of each (von Wright 1963; Doyle, Shoham, and Wellman 1991; Wellman and Doyle 1991; Hansson 1996; van der Torre 1997). It lets one state comparisons between both individuals and classes of individuals without trivializing tradeoffs by presupposing comparisons refer only to individuals that differ only in the indicated properties and that are indistinguishable with respect to all other properties. In its simplest form, this semantics also presupposes a decomposition of outcomes into attributes, as with the multiattribute utility, so that comparison *ceteris paribus* means comparison holding all attributes fixed except the ones involved in the preference statement. More refined semantics do not presuppose a decomposition into attributes, but instead directly presuppose the necessary relations of equivalence other things equal (Doyle and Wellman 1994). *Ceteris paribus* semantics implies independence relationships between qualities related to those implicit in multiattribute utility functions. The main languages studied so far lack a way of quantifying tradeoffs, and enforce a monotonicity that prevents specification of exceptions to preferences stated previously, although some work provides means for overriding general preferences with more specific ones (Loui 1990; Tan and Pearl 1994; van der Torre 1997).

Similarity-based semantics for preferences employs an idea of comparative similarity of outcomes from the semantics of conditionals, to restrict comparison to outcomes satisfying the indicated properties which are otherwise as similar as possible to each other (Boutilier 1994). Possible similarity notions include the *ceteris paribus* conception of being the same apart from the indicated properties, but also include broader notions of similarity. This semantics offers some of the same advantages as *ceteris paribus* semantics, and the additional advantage of permitting the statement of exceptions. Unfortunately, the general form of this semantics proves very weak, and supports almost no inference about preference.

Possibility-based semantics resemble similarity-based semantics, but offers somewhat stronger inferential relations (Dubois et al. 1998).

5.2. Nonmonotonic Preference Semantics

In comparison with logics of other concepts, exploration of possible semantics for preference statements remains in its infancy. Little reason exists to think these semantics exhaust the interesting classes of meanings conveyed or conveyable by people in preference expressions. Indeed, some fairly obvious combinations of semantics appear to offer some advantages, especially those involving extensions that encompass nonmonotonic exceptions.

For example, one might embed a monotonic *ceteris paribus* preference logic in a nonmonotonic logic. The monotonic preference logic of Doyle et al. (1991) already allows for specification of the preferences operating under certain conditions, as in

$$c_1 \implies (c_2 < c_3).$$

The nonmonotonic embedding would extend this to allow statement of preferences exceptional with respect to such operating conditions. For example, using the reason notation of Doyle (1988, 1994) (akin to the reasons of truth maintenance system (Doyle 1979) and the defaults of default logic (Reiter 1980)), one would write

$$c_1, \dots, c_n \parallel c'_1, \dots, c'_m \Vdash c'' < c'''$$

to indicate drawing the conclusion $c'' < c'''$ when one has each of the antecedent c_i conditions and none of the qualification or defeating c'_i conditions. Each nonmonotonic extension of axioms within the logic then represents a set of generic preferences; to get a consistent set of preferences the axioms must also enforce meanings matching the *ceteris paribus* semantics. At present we lack a logic that characterizes *ceteris paribus* semantics in terms of axioms that could be added to a standard nonmonotonic logic. In the absence of such axioms, one could obtain a nonmonotonic preference logic by embedding the semantical requirements in the definition of extensions, but this renders the definition of extensions quite complex even in the propositional case.

One clearly might apply the same nonmonotonic hybridization technique to the other semantics as well. The trick for each of these remains that of providing convenient expression over the semantical base.

The possibilities for hybrid preference logics multiply quickly as one contends with both variation in the interpretation of preference statements and variation in nonmonotonic statements. In particular, in addition to the variation in preference interpretations discussed in the preceding, one can interpret statements as involving either explicit or implicit nonmonotonicity. For example, the specificity-based overriding developed by Tan and Pearl (1994) constitutes a form of implicit nonmonotonicity, in that one need not defeat particular conditions to overturn some preference (see also van der Torre 1997). Further, one can understand either sort of nonmonotonicity in several ways, including in terms of exceptions that defeat derivations from current assumptions, in terms of statements violated only on a set of zero or infinitesimal probability, or in terms of statements that minimally deform the state of mind with respect to some notion of comparative similarity.

In fact, the possibility of nonmonotonic embeddings of preference logics raises new issues within preference logics themselves. One might consider modifying the strict *ceteris paribus* interpretation of Doyle et al. (1991) to instead employ some theory-dependent identification of which conditions can vary in the comparison. The sort of embedding just sketched might accomplish this itself. More interestingly, nonmonotonic reasons bear a natural interpretation as expressing preferences over mental states, so that the resulting states represent Pareto-optimal choices of state (see Doyle 1983b, 1985, 1994; Doyle and Wellman 1991; Horty 1994, or Delgrande et al. 2004 for a survey of recent means for incorporating preference information in nonmonotonic formalisms). Applying such an interpretation to nonmonotonic preference logic, one finds some statements of the theory expressing preferences over preferences in a manner akin to the preference-revision formalism discussed in Doyle (1990, 1992).

5.3. Other Extensions

Nonmonotonic extensions to preference languages start by following a pattern well-trodden in knowledge representation, but end up with new ideas about how to interpret preference statements. One might expect similar benefits from re-examination of other traditional representational ideas, including investigations of local, global, concrete, quantified, conditional, modal, and generic statements of preferences. Another natural direction would expand on extant graphical representations (e.g., Boutilier et al. 1999a) to incorporate some of these other traits of knowledge. In all these extensions, one wonders whether preferences exhibit any special structure that shapes the above aspects of their expression. Are there aspects of expression particular to preferences not listed above?

Many other questions remain unanswered. Should we be seeking to interpret the varied expressions of preference information in a single form of meaning? This certainly has the advantage of making it possible to compare the meanings of various expressions, but this

theoretical advantage may offer little practical help if the expressions take very different forms or prove so complex as to make the comparison even more complicated. How do the various expressions bias the specifications people might construct using them? How hard is it for each semantics to express notions from the others? How much do the meanings of standard expressions vary with semantics? To what extent do formal expressions mirror informal ones?

To what extent do the different representations fit different roles for preferences in preference acquisition, reasoning, decision modeling, decision making, maintenance, explanation, justification, summarization, and approximation? Can one formalize the requirements of these roles enough to automate choice of representation? To what extent can one translate between representations? Automatically? Accurately? Approximately? Efficiently?

6. STRUCTURE

Understanding the structure of economic preferences starts with understanding the structure of the corresponding preorders, or in the standard mathematical parlance, classifying all possible preorders. Mathematical classification proceeds by looking for formal concepts—logical, algebraic, geometric, or analytic—that provide leverage in characterizing the objects under study. For the case of preferences, mathematics provides numerous concepts of proven worth, with multiattribute decision theory, based on the algebra of functions on multidimensional spaces, providing the most familiar example.

Constraining preferences to have certain properties, whether from the nature of the agent exhibiting them, or from the origins of the preferences, can lead to a smaller and perhaps easier classification problem, but the great variety we see in people and in domains of action leads one to expect a difficult classification task whatever be the starting constraints. Indeed, one needs to consider additional constraints to obtain practicable representations. The theory of preordered sets is essentially identical to the theory of categories, an abstract mathematical notion that permeates modern developments of most mathematical subjects, and can serve, in special forms, as a foundation for all of mathematics.

One often first seeks to identify structure in the set Ω that makes utility functions U look simple, and then use the structure of the underlying set to inform the analysis of the preference relation. This approach has a long history in both mathematics and artificial intelligence. Boutilier et al. (1999b) present a good exposition related to decision theory.

Alternatively, one can first seek to identify structure in the preference relation to inform the analysis of the underlying set. Presumably this approach comes into play informally when decision analysts use their knowledge and intuition to guess at the sets of attributes appropriate for representing utility over outcomes. Formal methods within this approach have been pursued only in recent years (Holtzman 1989; Wellman, Breese, and Goldman 1992; Steward 1998). These formal methods include application of techniques aimed at detecting patterns in data sets and theoretical methods for inferring dimensional structure from piecemeal preference expressions.

In the long run, it should prove helpful to combine both these approaches, and look to iteratively identify improvements in understanding both preference and outcome structure.

6.1. Dimensional Structure

Looking at past practice in understanding preference relations and utility functions, we find the most common starting point to be the identification of dimensional structure in the

underlying set. Dimensional structure represents Ω in terms of the product $A = \prod_{i \in \mathcal{I}} A_i$ of sets A_i of attributes indexed by a set \mathcal{I} . Multiattribute decision theory reflects this approach directly by constructing utility functions over outcomes by composing simpler “subutility” functions defined over subsets of the attributes.

For example, one might characterize each alternative $\vec{x} = (x_1, \dots, x_n)$ in terms of five attributes and express the overall utility function in terms of functions over some of these attributes, such as

$$U(\vec{x}) = \alpha_{1,2}U_{1,2}(x_1, x_2) + \alpha_3U_3(x_3) + \alpha_{4,5}U_4(x_4)U_5(x_5), \quad (1)$$

where each U is a function over some set of attributes and each α is a numerical factor. The functional structure of such composite utility functions directly reflects the independence of the contributions to utility due to some attributes from the contributions due to others. One immediately sees in (1) that the contribution of the first two attributes stands separate from those of the other attributes. In some cases of subutility functions defined on overlapping sets of attributes, one can infer additional independencies from those manifest in the functional form (see Wellman 1985; Wellman and Doyle 1992). Several recent preference representations exploit explicit specification of independencies to state preference and utility information in simple forms (see Bacchus and Grove 1995, 1996; Shoham 1997a, 1997b; Boutilier et al. 1999a; La Mura and Shoham 1999).

6.1.1. Structure of Outcomes. One begins judging dimensional representations by examining the correspondences between the sets Ω and A , by examining whether each outcome and attribution corresponds (respectively) to some attribution or outcome, and whether these correspondences are one-to-one or many-to-one.

In a perfect attributive representation, each outcome corresponds to exactly one attribution, and vice versa, a circumstance we write as $\Omega \simeq A$. In the second best case, each outcome corresponds to exactly one attribution, but A contains elements that correspond to no outcome. In the third best case, every outcome corresponds to some attribution, but other outcomes might correspond to the same attribution. In the fourth and worst case, one lacks both unique and complete correspondence between outcomes and attributions.

If $\Omega \simeq A$, we say the dimensional representation is *exact*, as it permits an exact correspondence between outcomes and attributions. However, in seeking attributive representations one ordinarily rests content with attributions that uniquely identify outcomes. With this in mind, we say that a *framing* of a set of outcomes Ω is an injective (one-to-one) map $\phi : \Omega \rightarrow A$. One obtains framings from less accurate attributions by either expanding the space of attributes or attribute values, or by deciding to identify outcomes that map to the same attributes, so reducing the set of outcomes to fit the desired attributes.

6.1.2. Structure of Utility. Framings cleanly divide outcomes up into attributes that distinguish between the outcomes. Although such distinctive representations prove useful for many purposes, they need not best serve the analysis of the structure of utility. In understanding utility and preference, the primary focus rests on the structure of the utility function, not on the structure of the outcomes that possess utility. For this purpose, we seek attributions that cleanly divide up outcomes into attributes that correspond to distinct dimensions of variation of utility. That is, we seek coordinate systems over outcomes such that utility takes a simple form in the chosen coordinate system.

In the most pleasing situation, the right dimensional structure decomposes preference and utility into a simple form. This observation serves as the basis for multiattribute utility theory,

in which the right dimensional structure can yield a linear form for the utility function, with the utility function on the underlying set obtained as a weighted sum of subutility functions on the individual dimensions.

The main benefit of linear representations of this form is that they exhibit a maximal degree of utility or preferential independence. If the utility function has this form, we know that whenever two individuals differ only in one attribute, the overall preference between the two reflects the comparative preference or subutility along the dimension in which they differ.

Standard practice urges the identification of dimensions that permit linear forms for utility functions. In practice, the dimensions natural to human decision analysts provide partial but not full linearization, reducing the overall utility function to a weighted sum of subutility functions over subsets of dimensions, as in the form (1). One can sometimes reduce these subutility functions further to nonlinear combinations of sub-subutility functions, and thus obtain the overall utility as a multilevel linear and nonlinear composition of subutility functions (Keeney and Raiffa 1976; Wellman and Doyle 1992).

Can one do better? Can one recast the underlying set in terms of a different span of dimensions such that the utility function becomes linear? If so, can one find new linearizing dimensions that also mean something to human interpreters?

One trivial answer presents itself immediately. One can always find a linear representation of a utility function by viewing the underlying set as a single attribute and using the utility function as the sole subutility function. We look past this unifying answer by seeking transformations that preserve something of the dimensional structure, especially the linear structure already identified. Clearly any suitable transformation must be nonlinear. The transformation should preserve important topological and metric properties of the underlying set, unless the results remain intelligible without such properties.

The problem of finding linear representations of functions has motivated much of mathematical analysis, in which the notions of derivatives and tangent planes provide local linear representations. Given a function, analysis focuses on finding linear structure in the neighborhood of each point. Given local linear structure, analysis focuses on finding global functions that fit the local structure. Standard multiattribute utility theory employs these techniques, but typically focuses on functions for which the local linear approximations to the utility function hold globally, that is, in which the utility function has the same functional form at each point in the space of outcomes. Branting (1999) investigates utility representations that start by approximating local gradients in the utility function. Chajewska, Koller, and Parr (Chajewska and Koller 2000; Chajewska et al. 2000) exploit a different approach to locality by representing utility functions as piecewise probabilistic mixtures of prototypical utility functions.

Mathematics provides numerous conceptual tools for understanding the dimensional structure of utility and preference beyond the comparatively simple notions of partial derivatives and decomposition into subutility functions. These conceptual tools include locality analyses based on topologic and metric notions, and characterizations of global structure in terms of topological or differentiable manifolds. These global analyses explicitly aim to interpret complicated functions over simple spaces (e.g., a hyperplane) as simpler functions over more complex underlying spaces (e.g., curved and closed hypersurfaces).

Utility theory as developed by economists employs such more advanced notions, but many fundamental questions remain unanswered. For example, are there natural notions of “utility manifolds” that go beyond the sorts of manifolds common in mathematics and physics to incorporate constraints special to utility and preference? Would any global properties of utility manifolds reflect important facts about preference orders? Do the standard decompositions of functions in multiattribute utility theory exhaust these?

6.2. Determining Dimensional Structure

Supposing that the utility of interest has useful dimensional structure, say that presumed by multiattribute utility theory, just where do we get the outcome attributes or dimensions? Standard practice involves the hoary technique of the educated guess, in which the analyst uses experience to identify initial dimensions, and then corrects the guess as needed.

Artificial intelligence often starts with such manual modeling effort, but always then asks whether one can do better by mining important structure from data. In the case of preference and utility, automated analysis would seek to identify dimensional structure from preference expressions, from knowledge about the origins of preference, and from data about the behavior of the agent or agents in question. None of these approaches have been explored very thoroughly to date, and all would benefit from further investigation.

Perhaps the most extensive investigations have involved applying statistical techniques to the analysis of commercial data involving customer purchasing behavior or community recommendations. Generally the analysis starts with a set of dimensions (products, prices, rankings, etc.) given in advance by the application, and then tries to identify some similarities or structure in the data sets, as in Branting and Broos (1997) and Branting (1999, 2003). These analyses offer some results when the interesting dimensions of variation admit expression in terms of the initial set of dimensions, but yield less help when the structure of utility involves distinctions not captured in the ones assumed by the modeler.

Another means for identifying dimensional structure works by analyzing dependence relationships reflected in statements of preferences, such as statements in logics of preference *ceteris paribus*. Doyle et al. (1991), for example, interpret a statement that p is preferred to q as meaning that outcomes satisfying p are preferred to outcomes satisfying q , other things being equal, and interpret “other things being equal” in dimensional terms, using vectors of proposition values to represent outcomes, and comparing outcomes only when all dimensional values remain constant except those involved with p and q .

The first stage of analysis in this approach requires examination of logical relations among dimensions. One cannot vary logically dependent propositions independently, and so cannot use all propositions as independent attributes. One cannot, moreover, force informants to use only independent propositions, but instead must find ways of factoring out logical dependencies to better understand preferential dependencies. McGeachie (McGeachie 2002; McGeachie and Doyle 2002) approaches this task by examining the logical structure of a set of preference statements to distinguish the set of fundamental propositions involved in the preference statements from the set of fundamental propositions not involved in the preference statements.

The second stage of this analysis examines preferences relating logically independent propositions to identify preferential dependencies. As detailed by McGeachie (McGeachie 2002; McGeachie and Doyle 2002) and indicated by Boutilier et al. (1999a), this involves clustering propositions into preferentially dependent sets. One then proceeds to construct subutility functions for each subset of mutually dependent propositions, and then constructs overall utility functions as linear combinations of the independent subutility functions.

The logical formalization of the *ceteris paribus* approach is constrained to work with conditions expressible in terms of a stipulated propositional basis. This stipulated basis need not be the best one for illuminating the structure of preferences, thus one looks for methods that transform such stipulated coordinates to obtain a new basis of dimensions that reveal the independent aspects of preferential structure.

One approach to this problem of dimensional analysis, the contextual equivalence theory Doyle and Wellman (1994) avoids starting with linguistic expressions of preference that stipulate putatively “fundamental” propositions and instead constructs a basis of propositions

to fit a nonlinguistic or behavioral conception of “other things being equal”. This approach formalizes the treatment of outcomes as equivalent *ceteris paribus* with respect to two or more propositions as a function, called a *contextual equivalence*, from sets of propositions to equivalence relations over outcomes. Although both the domain and range of such a function form Boolean algebras, a general contextual equivalence need not form an algebraic homomorphism. A contextual equivalence that exhibits several natural algebraic and semantic properties, however, induces attribute representations that support propositional *ceteris paribus* reasoning. In this way, this approach constructs a dimensional “coordinate system” that exactly fits the identified behavioral equivalences.

6.3. Other Dimensional Analyses

What can be done starting with other structural concepts? In addition to the algebraic and geometric structures discussed in the preceding, mathematics provides for comparisons based on measure-theoretic, topological, and metric notions. One might interpret generic preferences in measure-theoretic terms by saying that one condition dominates another everywhere apart from a set of measure zero, or apart from a set of infinitesimal measure. One can use topological neighborhood notions to express nearness comparisons related to the comparative similarity relations of Lewis’s (1973) theory of counterfactuals. One can also employ similarity relations based on metrics or variants of the metric concept, including local and global metrics.

7. CHANGE

Can one capture preferences?

As noted previously, economic theory typically assumes that the preferences of an agent remain constant or change at a very slow rate compared to the agent’s beliefs. Is this assumption correct or reasonable? After all, reasoning and action changes one’s point of view (see Minsky 1980; Harman 1986), and one’s preferences constitute part of one’s point of view.

One can excuse the assumption of stable preferences by observing that one can push all the needed variation of preferences into the belief state of the agent, and then define a stable preference function that depends on the belief state in the appropriate way. This theoretical device deceives, however, as it requires either that one foresee all the possible ways preferences might evolve with time, or that one simply uses beliefs about preferences as an exact proxy for the preferences themselves. Experience shows that attaining such foresight is infeasible for any but the simplest agents performing fixed tasks in fixed circumstances; and without such fixity, even attempting to foresee all possible futures seems not worth the effort. The remaining alternative of using a special class of beliefs as a proxy for preferences merely renames change of preference as change of belief or state variable. This change of terminology says nothing about the correctness or reasonability of the economic assumption.

7.1. Sources of Change

From the viewpoint of artificial intelligence, preference change is pervasive and perpetual, occurring through experience, reasoning, and action.

Experience changes preferences routinely, through observation of one’s self and one’s environment. One discovers something is unexpectedly good or bad. One comes to like something more through familiarity, or to like something less.

Reasoning changes preferences in several ways. Planning routinely involves adopting or abandoning goals and subgoals, which in turn add desires and preferences. Adding a goal can mean adding a new dimension of evaluation. Fragmenting goals into subgoals adds new dimensions of preference; linking two goals means combining formerly independent dimensions into one. Deliberation over alternatives that offer new combinations of qualities involves developing new preferences that relate the alternatives at hand, whether by inference, induction, or invention (see, for example, Doyle 1980). Indeed, because the nonmonotonic reasons that pervade ordinary reasoning themselves embody preferential information in natural ways, as noted earlier, deliberative reasoning toward preferential conclusions produces reasons that express new preferences. Preference change also occurs as shifting alignments in the “society of mind” create new overall preferences (Hirschman 1982; Minsky 1986).

Action changes preferences by changing self and environment. Favoring one arm over the other increases its strength and skill in the activities for which it was favored; the same thing happens when repeated use of tools increases their fitness to the typical purpose. Repeated exposure to caffeine or other drugs changes bodily tolerances, leading to decreased preference for the same dose. At the mental level, action can involve self-construction from disparate reasons or considerations, such that the resulting construction embodies or includes new preferences (Doyle 1983b, 1989a).

7.2. Dealing with Change

Preference change engenders several problems for artificial intelligence and economic theory.

The first problem raised by preference change consists of changing the underlying structure of preferences from one well-represented by one set of independence relationships to a new structure that the old representation fits poorly. When do preference changes change underlying structure? When does learning of new constraints on the space of alternatives introduce new dependencies among dimensions of preferences, so that one cannot increase utility due to one attribute without decreasing utility due to another. How can one tell if preferences remain simple with respect to old dimensions or representational bases? What preference changes leave the set of independent dimensions invariant? When can one regain a set of independent dimensions by re-representation? Do some preference changes make such re-representation impossible? Do some preference changes indicate particular forms of representation change? That is, does knowledge of the specific change of preference aid in finding a new independent representation? If so, how do standard forms of experience, reasoning, and construction affect independence? Are there preference changes that induce linear changes in the independent dimensions?

The second problem raised by preference change involves how to track preference changes. If one has some model of preferences, how does one verify the model and its continuing correctness? As in other verification tasks, one can seek to verify correctness observationally, whether in the course of events or through deliberate sampling and directed experiment. When observations suggest incorrectness, how should one diagnose changes and assign credit or blame to underlying preferences, plans, or beliefs? The monitoring and maintenance needed to cope with pervasive preference change motivates integration of ongoing preference learning and adaptation into reasoning and action. We do not yet understand the ramifications of the tracking task, especially when the very act of learning or observation can change preferences. Is preference “elicitation” a fantasy? Can one tell whether acquisition has changed preferences? Can one predict the types of changes produced by interactions or diagnose the actual consequences?

The third problem of preference change is how to revise preferences. One always looks for revision processes to exhibit some degree of rationality where possible. The Bayesian mixture approach (Chajewska and Koller 2000; Chajewska et al. 2000) offers one rational process for revising approximations to representing utility functions, but can fail dramatically when the target utility function changes, and so cannot be used to track changing preferences without substantial modification. In the case of preference-guided revision of preference, the aim of rational revision raises issues of whether one judges the rationality of changes prospectively by the preferences obtaining before the change or retrospectively by the preferences obtaining after the change. Jeffrey (1983) introduced the term “ratified rational choice” for retrospectively rational decisions. Doyle (1990, 1991) examined a formulation of rational revision processes that applied to both belief and preference, as well as other mental attitudes. This formulation reinterprets Gärdenfors–Makinson (Gärdenfors and Makinson 1988) entrenchment orders as indicating preference orders over mental states, and formalized rational revision both in terms of prospective rational choices and in terms of retrospective “ratichoice” revision, as well as some nonmonotonic variants. Rational revision can violate some of the revision postulates proposed by Alchourrón, Gärdenfors, and Makinson (1985), but the remaining principles of revision, such as they are, do not provide much predictive power regarding preference change. More to the point, because the preferences about revisions enter into their own revision, the problem becomes one of adjudicating among competing preference criteria of specificity, probability, authority, and seniority or recency. This adjudication process endows the preference revision problem with the character of a group decision process, and so raises the difficulties exemplified by Arrow’s impossibility theorem (Arrow 1963; Doyle and Wellman 1991).

8. CONCLUSIONS

Preference information plays a crucial role in understanding and controlling effective and adaptive reasoning and action. We have understood some straightforward quantitative and qualitative representations and semantics, and have simple and useful examples of their application in automated systems. These initial explorations make clear that the full concept of preference involves more than mere ordering, for the preference ordering must bear some sensible relation to actions of the agent, at least in light of the beliefs, intentions, and other attitudes of the agent.

However, we have only begun to understand the variety, nature, and representation of preferences, or the appropriate methods for specifying, realizing, effecting, tracking, adapting, and revising preferences. Effective acquisition and use of preference information requires understanding the instantaneous geometric structure of preference and, because preference change is pervasive, the evolving geometric structure of preference. The origins of preference shape this geometry; constraints on change shape the evolution of this geometry; and the purposes of reasoning and action shape effective representations of this geometry and utility functions over it.

ACKNOWLEDGMENTS

This article grew out of an invited presentation at the 2002 AAI Workshop on *Preferences in AI and Constraint Programming: Symbolic Approaches*. I thank the organizers for the invitation to present these ideas in that forum and their encouragement to write them down. I have been helped in this writing by discussions with Michael McGeachie, Robert Laddaga,

Peter Szolovits, and Howard Shrobe. I am especially grateful to James Delgrande for careful reading that improved the paper, and to McGeachie for other valuable comments on drafts.

REFERENCES

- ALCHOURRÓN, C. E., P. GÄRDENFORS, and D. MAKINSON. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, **50**:510–530.
- ARROW, K. J., and H. RAYNAUD. 1986. *Social Choice and Multicriterion Decision-Making*. MIT Press, Cambridge, MA.
- ARROW, K. J. 1963. *Social Choice and Individual Values* (2nd ed.). Yale University Press, New Haven.
- BACCHUS, F., and A. GROVE. 1995. Graphical models for preference and utility. *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pp. 3–19, San Francisco, CA.
- BACCHUS, F., and A. GROVE. 1996. Utility independence in a qualitative decision theory. *In Proceedings of the Fifth International Conference on Knowledge Representation and Reasoning*, Morgan Kaufmann, pp. 542–552, San Francisco, CA.
- BOLLES, R. C. 1975. *Theory of Motivation* (2nd ed.). Harper and Row, New York.
- BOUTILIER, C. 1994. Toward a logic for qualitative decision theory. *In Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR'94)*. Edited by J. Doyle, E. Sandewall, and P. Torasso. Morgan Kaufmann, San Francisco, CA.
- BOUTILIER, C., R. I. BRAFMAN, H. H. HOOS, and D. POOLE. 1999a. Reasoning with conditional *ceteris paribus* preference statements. *In Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, Stockholm, pp. 71–80. Edited by Henri Prade, Kathryn Laskey. Published by Morgan Kaufmann, San Francisco, CA.
- BOUTILIER, C., T. DEAN, and S. HANKS. 1999b. Decision theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, **11**:1–94.
- BRANTING, L. K. 1999. Active exploration in instance-based preference modeling. *In Proceedings of the Third International Conference on Case-Based Reasoning (ICCB-99)* Monastery Seon, Germany.
- BRANTING, L. K. 2003. Learning feature weights from customer return-set selections. *Knowledge and Information Systems*, L. Karl Branting, Learning Feature Weights from Customer Return-Set Selections, *The Journal of Knowledge and Information Systems (KAIS)* **6**(2) March (2004).
- BRANTING, L. K., and P. BROOS. 1997. Automated acquisition of user preferences. *International Journal of Human-Computer Studies*, **46**:55–77.
- CHAJEWSKA, U., and D. KOLLER. 2000. Utilities as random variables: Density estimation and structure discovery. *In Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI-2000)*, AUAI.
- CHAJEWSKA, U., D. KOLLER, and R. PARR. 2000. Making rational decisions using adaptive utility elicitation. *In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-00)*, AAAI Press, AAAI, pp. 363–369.
- de KLEER, J., J. DOYLE, G. L. STEELE, JR., and G. J. SUSSMAN. 1977. AMORD: Explicit control of reasoning. *In Proceedings of the ACM Symposium on Artificial Intelligence and Programming Languages*, pp. 116–125, Rochester, New York.
- DELGRANDE, J., T. SCHAUB, H. TOMPITS, and K. WANG. 2004. A classification and survey of preference handling approaches in nonmonotonic reasoning. *Computational Intelligence*, **20**(2): 308–334.
- DOYLE, J. 1979. A truth maintenance system. *Artificial Intelligence*, **12**(2):231–272.
- DOYLE, J. 1980. A model for deliberation, action, and introspection. AI-TR 581, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA, 02139.
- DOYLE, J. 1982. The foundations of psychology. Technical Report 82-149, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

- DOYLE, J. 1983a. Methodological simplicity in expert system construction: The case of judgments and reasoned assumptions. *AI Magazine*, 3(2):39–43.
- DOYLE, J. 1983b. Some theories of reasoned assumptions: An essay in rational psychology. Technical Report 83-125, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA.
- DOYLE, J. 1985. Reasoned assumptions and Pareto optimality. *In Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pp. 87–90, Los Angeles, CA.
- DOYLE, J. 1988. Artificial intelligence and rational self-government. Technical Report CS-88-124, Carnegie-Mellon University Computer Science Department.
- DOYLE, J. 1989a. Constructive belief and rational representation. *Computational Intelligence*, 5(1):1–11.
- DOYLE, J. 1989b. Reasoning, representation, and rational self-government. *In Methodologies for Intelligent Systems*, Vol. 4. *Edited by* Z. W. Ras. North-Holland, New York, pp. 367–380.
- DOYLE, J. 1990. Rational belief revision. Presented at the Third International Workshop on Nonmonotonic Reasoning, Stanford Sierra Camp, CA, Available at <http://www.csc.ncsu.edu/faculty/doyle/>.
- DOYLE, J. 1991. Rational belief revision (preliminary report). *In Proceedings of the Second Conference on Principles of Knowledge Representation and Reasoning*. *Edited by* R. E. Fikes, and E. Sandewall. Morgan Kaufmann, San Mateo, CA, pp. 163–174.
- DOYLE, J. 1992. Reason maintenance and belief revision: Foundations vs. coherence theories. *In Belief Revision*. *Edited by* P. Gärdenfors. Cambridge University Press, Cambridge, MA, pp. 29–51.
- DOYLE, J. 1994. Reasoned assumptions and rational psychology. *Fundamenta Informaticae*, 20(1–3):35–73.
- DOYLE, J., and M. MCGEACHIE. 2003. Exercising qualitative control in autonomous adaptive survivable systems. *In Revised papers from the Second International Workshop on Self-Adaptive Software (IWSAS 2)*, May 2001, Springer-Verlag, Berlin, pp. 158–170.
- DOYLE, J., Y. SHOHAM, and M. P. WELLMAN. 1991. A logic of relative desire (preliminary report). *In Methodologies for Intelligent Systems*, 6, Vol. 542 of Lecture Notes in Artificial Intelligence. *Edited by* Z. W. Ras and M. Zemankova. Springer-Verlag, Berlin, pp. 16–31.
- DOYLE, J., and R. H. THOMASON. 1999. Background to qualitative decision theory. *AI Magazine*, 20(2):55–68.
- DOYLE, J., and M. P. WELLMAN. 1991. Impediments to universal preference-based default theories. *Artificial Intelligence*, 49(1–3):97–128.
- DOYLE, J., and M. P. WELLMAN. 1994. Representing preferences as *ceteris paribus* comparatives. *In Proceedings of the AAAI Spring Symposium on Decision-Theoretic Planning*. *Edited by* S. Hanks, S. Russell, and M. P. Wellman, Stanford, CA.
- DUBOIS, D., D. LE BERRE, H. PRADE, and R. SABBADIN. 1998. Logical representation and computation of optimal decisions in a qualitative setting. *In Proceedings of AAAI-98*, AAAI Press, Menlo Park, CA, pp. 588–593.
- FRANKFURT, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy*, 68:5–20.
- GÄRDENFORS, P., and D. MAKINSON. 1988. Revisions of knowledge systems using epistemic entrenchment. *In Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*. *Edited by* M. Y. Vardi. Morgan Kaufmann, Los Altos, CA, pp. 83–95.
- GÄRDENFORS, P., and N.-E. SAHLIN (Eds.). 1988. *Decision, Probability, and Utility: Selected Readings*. Cambridge University Press, Cambridge, MA.
- GORDON, T. F. 1995. *The Pleadings Game: An Artificial Intelligence Model of Procedural Justice*. Kluwer, Dordrecht.
- HADDAWY, P., and S. HANKS. 1998. Utility models for goal-directed decision-theoretic planners. *Computational Intelligence*, 14(3):392–429.
- HANSSON, S. O. 1996. What is *Ceteris Paribus* preference? *Journal of Philosophical Logic*, 25(3):307–332.
- HARMAN, G. 1986. *Change in View: Principles of Reasoning*. MIT Press, Cambridge, MA.
- HIRSCHMAN, A. O. 1982. *Shifting Involvements: Private Interest and Public Action*. Princeton University Press, Princeton.

- HOLTZMAN, S. 1989. *Intelligent Decision Systems*. Addison-Wesley, Boston, MA.
- HORTY, J. F. 1994. Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, **23**:35–65.
- JAMES, W. 1890. *The Principles of Psychology*. Henry Holt & Co., New York. (In two volumes.)
- JAMES, W. 1897. *The Will to Believe and Other Essays in Popular Philosophy*. Longmans, Green, and Co., New York.
- JEFFREY, R. C. 1983. *The Logic of Decision* (2nd ed.). University of Chicago Press, Chicago.
- KAHNEMAN, D., P. SLOVIC, and A. TVERSKY (Eds.). 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK.
- KEENEY, R. L., and H. RAIFFA. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley and Sons, New York.
- LA MURA, P., and Y. SHOHAM. 1999. Expected utility networks. *In Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pp. 366–373. Stockholm, pp. 71–80. *Edited by* Henri Prade, Kathryn Laskey, and Morgan Kaufmann, San Francisco, CA.
- LEWIS, D. 1973. *Counterfactuals*. Blackwell, Oxford.
- LOUI, R. 1990. Defeasible specification of utilities. *In Knowledge Representation and Defeasible Reasoning*. *Edited by* H. E. Kyburg, Jr., R. P. Loui, and G. N. Carlson. Kluwer Academic Publishers, Dordrecht, pp. 345–359.
- LUCE, R. D., and H. RAIFFA. 1957. *Games and Decisions*. Wiley, New York.
- MACHINA, M. J. 1987. Choice under uncertainty: Problems solved and unsolved. *Journal of Economic Perspectives*, **1**(1):121–154.
- MCCARTHY, J. 1979. Ascribing mental qualities to machines. *In Philosophical Perspectives in Artificial Intelligence*. *Edited by* M. Ringle, Harvester Press, Brighton, pp. 161–195.
- MCDERMOTT, D. 1978. Planning and acting. *Cognitive Science*, **2**:71–109.
- MCGEACHIE, M. 2002. Utility functions for *ceteris paribus* preferences. Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA.
- MCGEACHIE, M., and J. DOYLE. 2002. Efficient utility functions for *ceteris paribus* preferences. *In Proceedings of the AAAI Eighteenth National Conference on Artificial Intelligence (AAAI-2002)*. *Edited by* Rina Dechter, Michael Kearns, and Rich Sutton, Edmonton, Alberta, pp. 279–284.
- MINSKY, M. 1980. K-lines: A theory of memory. *Cognitive Science*, **4**:117–133.
- MINSKY, M. 1986. *The Society of Mind*. Simon and Schuster, New York.
- NAGEL, T. 1979. The fragmentation of value. *In Mortal Questions*, Chapter 9. Cambridge University Press, Cambridge.
- PASCAL, B. 1962. *Pensées sur la religion et sur quelques autres sujets*. Harvill, London. Translated by M. Turnell, originally published in 1662.
- RAMSEY, F. P. 1931. Theories. *In The Foundations of Mathematics and other Logical Essays*. *Edited by* R. B. Braithwaite. Routledge and Keegan Paul, London.
- RAO, A. S., and M. P. GEORGEFF. 1991. Modeling rational agents within a BDI-architecture. *In Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*. *Edited by* J. Allen, R. Fikes, and E. Sandewall. Morgan Kaufmann publishers Inc., San Mateo, CA, pp. 473–484.
- REITER, R. 1980. A logic for default reasoning. *Artificial Intelligence*, **13**:81–132.
- RUSSELL, S. J., and P. NORVIG. 2002. *Artificial Intelligence: A Modern Approach* (2nd ed.). Prentice-Hall, Englewood Cliffs, NJ.
- SHAND, A. F. 1920. *The Foundations of Character: Being a Study of the Tendencies of the Emotions and Sentiments* (2nd ed.). Macmillan and Company, London.
- SHOHAM, Y. 1987. Nonmonotonic logics: Meaning and utility. *In Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, Milan, Italy, pp. 388–393.

- SHOHAM, Y. 1997a. Conditional utility, utility independence, and utility networks. *In Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence. Edited by D. Geiger, and P. P. Shenoy.* Morgan Kaufmann, San Francisco, CA, pp. 429–436.
- SHOHAM, Y. 1997b. A symmetric view of probabilities and utilities. *In Proceedings of IJCAI-97. Edited by M. E. Pollack.* Morgan Kaufmann, San Francisco, CA, pp. 1324–1329.
- STEWART, D. A. 1998. Utility assessment based on individualized patient perspectives. Ph. D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 02139.
- TAN, S.-W., and J. PEARL. 1994. Qualitative decision theory. *In Proceedings of the Twelfth National Conference on Artificial Intelligence, AAAI Press, Menlo Park, CA.*
- VAN DER TORRE, L. 1997. Reasoning about obligations: Defeasibility in preference-based deontic logic. Tinbergen Institute research series, Erasmus University, Rotterdam.
- VAN FRAASSEN, B. C. 1973. Values and the heart's command. *Journal of Philosophy*, **LXX**(1):5–19.
- VON WRIGHT, G. H. 1963. *The Logic of Preference: An Essay.* Edinburgh University Press, Edinburgh.
- WELLMAN, M. P. 1985. Reasoning about preference models. TR 340, Massachusetts Institute of Technology, Laboratory for Computer Science, Cambridge, MA, 02139.
- WELLMAN, M. P., J. S. BREESE, and R. P. GOLDMAN. 1992. From knowledge bases to decision models. *The Knowledge Engineering Review*, **7**(1):35–53.
- WELLMAN, M. P., and J. DOYLE. 1991. Preferential semantics for goals. *In Proceedings of the Ninth National Conference on Artificial Intelligence, Anaheim, CA, pp. 698–703.*
- WELLMAN, M. P., and J. DOYLE. 1992. Modular utility representation for decision-theoretic planning. *In Proceedings of the First International Conference on AI Planning Systems, College Park, MD, pp. 236–242.*