Reprinted from *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, July, 1991, pp. 698–703.

Preferential Semantics for Goals

Michael P. Wellman
Wright Laboratory AI Office
WL/AAA-1
Wright-Patterson AFB, OH 45433
wellman@wrdc.af.mil

Jon Doyle*
MIT Lab for Computer Science
545 Technology Square
Cambridge, MA 02139
doyle@zermatt.lcs.mit.edu

Abstract

Goals, as typically conceived in AI planning, provide an insufficient basis for choice of action, and hence are deficient as the sole expression of an agent's objectives. Decision-theoretic utilities offer a more adequate basis, yet lack many of the computational advantages of goals. We provide a preferential semantics for goals that grounds them in decision theory and preserves the validity of some, but not all, common goal operations performed in planning. This semantic account provides a criterion for verifying the design of goal-based planning strategies, thus providing a new framework for knowledge-level analysis of planning systems.

Planning to achieve goals

In the predominant AI planning paradigm, planners construct plans designed to produce states satisfying particular conditions called *goals*. Each goal represents a partition of possible states of the world into those satisfying and those not satisfying the goal. Though planners use goals to guide their reasoning, the crude binary distinctions defined by goals provide no basis for choosing among alternative plans that ensure achievement of goals, and no guidance whatever when no such plans can be found. These lacunae pose significant problems for planning in all realistic situations, where actions have uncertain effects or objectives can be partially satisfied.

To overcome these widely-recognized expressive limitations of goals, many AI planners make *ad hoc* use of heuristic evaluation functions. These augment the guidance provided by goals, but lack the semantic justification needed to evaluate their true efficacy. We believe that heuristic evaluation functions should not be viewed as mere second-order refinements on the primary goal-based representation of objectives, supporting a separate "optimizing" phase of planning. Our thesis is that relative preference over the possible results of a plan constitutes the fundamental concept underlying the objectives of planning, with goals serving as a computationally useful heuristic approximation to these preferences (Doyle, 1990). Our purpose here is to provide a formal semantics for goals in terms of decision-theoretic preferences that supports rational justifications for planning principles. The grounding in decision theory enables designers to determine whether their planning systems act rationally in accord with their goals, and provides a principled basis for integrating goals with other types of preference information.

We begin by summarizing some basic concepts of preference. We then develop formal decision-theoretic semantics for goals and examine some standard planning operations in light of the semantics. We conclude by discussing some related work and offering some directions for future investigation.

Preferences and utility

Decision theory starts with the notion of *preferences* over *outcomes* (Keeney and Raiffa, 1976; Savage, 1972). Outcomes represent the possible consequences of the agent's decisions. In the planning context, an outcome might be taken to be the state resulting from execution of a plan, or perhaps the entire history of instantaneous states over the lifetime of the agent. To provide an adequate basis for decision, the set Ω of possible outcomes must distinguish all consequences that the agent cares about and are possibly affected by its actions. We represent the agent's preferences by a total preorder (a complete, reflexive, and transitive relation) $\succsim$ over Ω, called the *preference order*. When $\omega \succsim \omega'$ we say that ω is *weakly preferred* to ω', which means that the former outcome is at least as desirable as the latter. The strict preference order $\succ$ consists of the irreflexive part of $\succsim$, that is, $\omega \succ \omega'$ (ω is *preferred* to ω') if and only if (iff) $\omega \succsim \omega'$ but $\omega' \not\succsim \omega$. When both $\omega \succsim \omega'$ and $\omega' \succsim \omega$, we say the two outcomes are *indifferent*, and write $\omega \sim \omega'$. Decision theory postulates that rational agents make choices so that the chosen alternatives are maximally preferred among those available. In planning, agents choose among courses of action, or *plans*.

In a perfectly predictable or *deterministic* environment, the situation in which a plan is applied uniquely

---

*Jon Doyle is supported by the USAF Rome Laboratory and DARPA under contract F30602-91-C-0018.

determines the outcome. For each situation s, we write $\rho_s(\pi)$ to denote the result of executing the plan π in s. When the situation is fixed or clear from context, as in choosing among plans from a particular state, we omit the situation subscript and just write ρ. Under conditions of determinism, preferences on outcomes induce preferences on plans: $\pi \succsim \pi'$ iff $\rho(\pi) \succsim \rho(\pi')$. In the more common case of *uncertainty*, plans influence outcomes only probabilistically, and we must replace ρ by a probability distribution over Ω (called a *prospect*), conditional on π. Accounting for uncertainty requires that we enlarge the domain of $\succsim$ to express preferences over the set of all prospects.

Much of decision theory is concerned with conditions under which $\succsim$ is representable by an order-preserving, real-valued *utility function*, and with identifying regularities in preferences that justify utility functions with convenient structural properties (Keeney and Raiffa, 1976). Although we expect that utility theory will have much to offer for the design of planning systems, our basic preferential semantics for goals depends only on the underlying preference relation.

What's in a goal?

What does it mean to say that an agent has a goal γ? The most direct interpretation would define the problem to have two relevant outcomes, $\Omega = \{\gamma, \neg\gamma\}$, with a preference order consisting of $\gamma \succ \neg\gamma$. Any two-valued utility function u satisfying $u(\gamma) > u(\neg\gamma)$ would suffice to represent this preference. However, this simple preferential interpretation of goals is inadequate for several reasons.

First, goals serve a dual role in most planning systems, capturing aspects of both *intentions* and *desires* (Doyle, 1980). Besides expressing the desirability of a state, adopting a goal represents some commitment to pursuing that state. These two attitudes must be disentangled in any semantical treatment of goals. In our treatment, we concentrate exclusively on the role of expressing desirability, recognizing that the result is only a partial account of the use of goals in planning systems. For an analysis focusing on intentions, see, for example, Cohen and Levesque (1990). Ultimately we seek a comprehensive theory of goals addressing both their preferential and intentional facets.

Second, even if we limit our attention to desirability, the two-outcome interpretation described above falls short because it considers only a single goal. In particular, it says nothing about the important issues of how planners should combine, specialize, reformulate, trade off, or otherwise manipulate sets of goals.

Finally, the binary utility function interpretation provides a questionable basis for analyzing planning systems because decisions concerning a particular goal ordinarily have consequences for other factors that the agent cares about. The set Ω must thus include more than two possible outcomes to differentiate all the relevant factors, in which case the single-goal outcomes γ and $\neg\gamma$ correspond to a *partition* $\{\hat{\gamma}, \neg\hat{\gamma}\}$ of Ω. But the binary preference interpretation fails in this setting, as $\hat{\gamma}$ and $\neg\hat{\gamma}$ are sets of possible outcomes, not individuals, and preferences are variable within each set. Consider, for example, the familiar "hungry monkey." Outcomes satisfying the goal `has-bananas` might differ on how long it took to get the bananas, the quality or quantity of bananas possessed, whether the monkey slips on a peel along the way, or whether it wins the state lottery ten years later. These finer distinctions mean that many outcomes satisfying `has-bananas` are actually less desirable than many others satisfying ¬`has-bananas`, and plans attending only to `has-bananas` likely result in significant suboptimality. Yet analyses taking goals as the gold standard for preference would reveal no lack of rationality in the monkey's behavior.

We maintain that a diversity of relevant objectives characterizes most, if not all, planning situations. Even in situations where two values seem sufficient to describe the final outcome, as in chess, it often appears necessary for control of search to evaluate medial situations (*e.g.,* board positions) in terms of intermediate utilities. In such cases, the appropriate utility measure is the probability of goal achievement (Good, 1962). But even in chess, winning isn't everything, and the two-outcome model merely approximates the more precise preference structure that would consider the possibility of draws and the relative importance of games within a broader match or tournament context.

In the following sections, we answer the question "what's in a goal?" by providing an interpretation of goal predicates in terms of preference orders. We then consider the constraints induced by goals on rationally chosen plans.

Preferential semantics

A *proposition* is a subset of the outcome space, Ω, *i.e.,* the set of outcomes where the proposition holds. A *goal proposition* is, intuitively, a proposition comprising some desirable property of the outcome. We formalize this intuition by specifying conditions on the preference order under which a given proposition can be termed a goal.

The underlying idea of our semantics is that each goal proposition determines a "dimension" along which outcomes may vary. We view the goal proposition as indicating a direction along this dimension, with its complementary proposition indicating the opposite direction. We then call a proposition a goal just in case utility increases in the direction defined by the proposition, *ceteris paribus* ("all else being equal"). This approach suggests a direct generalization to multivalent, or non-propositional, outcome features influencing preference (such as the cost of some activity), as long as each feature can be ordered in the direction of increasing preference.

Since the set of outcomes lacks any particular structure at this point, the primary effort of formalizing the intuitive semantics lies in providing ways of determining directions and when "other things" are equal. In this paper we follow the path of analytic geometry and multiattribute utility theory and factor the outcome space into the cartesian product of a number of smaller spaces. The factor spaces correspond to dimensions or "attributes," and "all else being equal" means varying one attribute while holding all others constant. Elsewhere, we present a more general formalization that avoids coordinate systems in favor of a "metric" over outcomes closely related to standard theories of counterfactuals (Wellman *et al.*, 1991).

Definition 1 (Framings) *A* framing *of a set of outcomes Ω is an injective (one-to-one) map $\phi : \Omega \to A$, where $A = \prod_{i=1}^{n} A_i$.*

A framing ϕ of Ω induces an isomorphism of Ω with its image $\phi(\Omega) \subseteq A$, and we call ϕ *exact* just in case it indicates an isomorphism $\Omega \cong A$ with all of A. Because framings are one-to-one, they directly represent the chosen outcome space without blurring distinctions or introducing new ones. That is, they just provide alternative names for existing outcomes. The cartesian structure of the attributes space induces projection functions $\phi_i : \Omega \to A_i$ such that $\phi(\omega) = \langle \phi_1(\omega), \ldots, \phi_n(\omega) \rangle$ for each $\omega \in \Omega$. For notational convenience, we sometimes identify an outcome with its representation and assume that preferences over representations mirror preferences over the outcomes they represent. We abbreviate projections by subscripts so that, for example, ω_i means $\phi_i(\omega)$.

One ordinarily introduces framings in order to pick out certain attribute values as targets or thresholds. For example, the proposition $\{\omega \in \Omega \mid \omega_i = a\}$ consists of all outcomes for which the ith attribute achieves the target value a. Similarly, if A_i is ordered by a relation $\sqsubseteq_i$, then the proposition $\{\omega \in \Omega \mid a \sqsubseteq_i \omega_i\}$ consists of all outcomes for which the ith attribute value meets or exceeds the threshold value a.

The most interesting framings pick out significant dimensions along which utility varies. We say that a framing is *redundant* just in case there is some dimension i that is completely determined by the other dimensions, or formally, that for all $\omega \in \Omega$, we have $x = y$ whenever $x_j = y_j = \omega_j$ for all $j \neq i$. Similarly, a framing is *preferentially redundant* just in case there is a dimension i that is neutral with respect to preference, *i.e.,* for all $\omega \in \Omega$, $x \sim y$ whenever $x_j = y_j = \omega_j$ for all $j \neq i$.

We define goals relative to framings that distinguish the goal proposition as an attribute. Let γ be a proposition and $\phi = \langle \alpha, \beta \rangle$ a framing of Ω, where $\alpha : \Omega \to \{\gamma, \neg\gamma\}$.

Definition 2 (Goal) *γ is a* goal *in ϕ, written goal(γ, ϕ), just in case for all $\omega \in \Omega$, $(\gamma, \omega_\beta) \succsim (\neg\gamma, \omega_\beta)$ whenever both (γ, ω_β) and $(\neg\gamma, \omega_\beta)$ are in $\phi(\Omega)$.*

Definition 3 (Strict Goal) *γ is a* strict goal *in ϕ, written GOAL(γ, ϕ), just in case goal(γ, ϕ) and $\neg$goal$(\neg\gamma, \phi)$.*

According to these definitions, γ is a goal just in case any outcome in γ is weakly preferred to its corresponding outcome—if any—in $\neg\gamma$, holding constant the *residual* attributes given by β. It is a strict goal when at least one of these preferences is strict. Residual factors may sometimes render a goal irrelevant (through indifference between outcomes in γ and $\neg\gamma$) but cannot cause a preference reversal with respect to the goal. Moreover, a strict goal cannot be entirely irrelevant because preference is strict for at least one value of the residual. The *ceteris paribus* condition that outcomes be compared with respect to fixed values of the residual serves two purposes. First, the reference to context allows us to avoid the unrealistic assertion that any outcome achieving the goal is preferred to any that does not. And second, by quantifying over these contexts, we are permitted to compare preferences in particular situations, where something is known about the values of residual outcome attributes.

Finally, we note that this definition covers the fully multiattribute case since using the residual attribute β to represent several attributes (with $\beta = \langle \phi_2, \ldots, \phi_n \rangle$) requires no substantial change to the definitions.

The relativity of goals

Goalhood of a proposition depends in general on the framing of the outcome space. For example, consider an outcome space Ω consisting of all combinations of three logically independent propositions: p, "I am wearing a raincoat"; q, "I am out in the rain"; and r, "My dog has no fleas." We assume that r is preferred to $\neg r$, all else equal, that $\neg q$ is preferred to q, all else equal, and that p is preferred to $\neg p$ given q, but the preference is reversed given $\neg q$, again all else equal. The exact framing with attributes corresponding to each of these propositions yields the intuitive results that $\neg q$ and r are goals, but neither p nor $\neg p$ is a goal. Yet p is a goal in the (nonredundant) framing of Ω with attributes p, r, $p \wedge \neg q$, and $\neg p \wedge \neg q$, because we cannot vary p in this framing—holding all else equal—unless q also holds. Hence, the only situations which can be compared are those where p is preferred to $\neg p$.

However, goals need not depend entirely on particular choices of framings. In fact, suitably related framings support related goals, and some goals do not depend at all on how one represents residuals. Let $\phi = \langle \alpha, \beta \rangle$ and $\phi' = \langle \alpha, \beta' \rangle$ be alternative framings of Ω with $\alpha : \Omega \to \{\gamma, \neg\gamma\}$, $\beta : \Omega \to B$, and $\beta' : \Omega \to B'$. We say that ϕ *subsumes* ϕ' iff there exists a mapping $f : B' \to B$ such that ${\beta'}^{-1}(b) \subseteq \beta^{-1}(f(b))$ for all $b \in B'$. In other words, for every residual proposition expressible in ϕ', there is a corresponding residual in ϕ that includes a superset of its outcomes. Note that every framing subsumes itself and that exact framings of Ω subsume all other framings of Ω.

Proposition 1 *If ϕ subsumes ϕ', then*

1. *$goal(\gamma, \phi)$ implies $goal(\gamma, \phi')$, and*
2. *$GOAL(\gamma, \phi')$ implies $\neg goal(\neg\gamma, \phi)$.*

Thus goalhood in an exact framing implies goalhood in all framings. However, ambiguity in an exact framing admits a strong frame dependence.

Proposition 2 *If there is some framing $\langle \alpha, \beta \rangle$ of Ω, $\alpha : \Omega \to \{\gamma, \neg\gamma\}$, such that neither γ nor $\neg\gamma$ is a strict goal, then either γ is preference neutral in every framing of Ω, or there exist nonredundant framings $\phi' = \langle \alpha, \beta' \rangle$ and $\phi'' = \langle \alpha, \beta'' \rangle$ such that $GOAL(\gamma, \phi')$ and $GOAL(\neg\gamma, \phi'')$.*

The *ceteris paribus* condition of Definition 2 is a form of what multiattribute utility theory calls *preferential independence* (Gorman, 1968; Keeney and Raiffa, 1976), which requires that preference for each goal attribute be independent of the other attributes. (The usual definition of preferential independence, however, does not allow strictness to vary as in the goal definitions.) When one set of attributes does not exhibit preferential independence, we can sometimes restructure the outcome space into attributes that do. For example, one may incorporate the necessary qualifications into the goal proposition. In our example above, reframing with attributes $\neg q \vee p$, $\neg p \vee q$, q, and r renders the proposition $\neg q \vee p$ a goal. Alternatively, one may express the goal in terms of more fundamental attributes (Keeney, 1981). For example, the goal "I am dry" is a deeper expression of our preference for wearing raincoats when out in the rain. Taking yet another approach, we may express the goal conditionally, that is, with respect to a framing of a subset of the outcomes. In the example, p is a goal in an appropriate exact framing of the reduced outcome set $\Omega' = q \subseteq \Omega$. Though straightforward, a comprehensive treatment of these approaches falls beyond the scope of this paper.

Finally, we note that achieving a goal does *not* imply an improvement in expected utility because while the goal is preferred to its contrary *ceteris paribus*, it may have negative consequences via its probabilistic and logical relation to other attributes.

Preferences from goals

Definition 2 shows how to define goals in terms of preferences. In this section, we show, conversely, how to derive preferences from sets of goals, and discuss the implications of these preferences for choices of plans.

Each goal proposition constrains the preference order over Ω; combining several goals yields a partial specification of the complete order, with preference between competing goals or alternate ways of achieving the same goal not defined. Let $\Gamma = \{\gamma_1, \ldots, \gamma_m\}$ be a set of goals in a framing $\phi = \langle \alpha_1, \ldots, \alpha_m, \beta \rangle$ such that $\alpha_i : \Omega \to \{\gamma_i, \neg\gamma_i\}$ for each $i = 1, \ldots, m$.

Definition 4 (Goal Preferences) *We say that outcome ω is* goal-preferred *to ω' with respect to Γ in ϕ, written $\omega \succsim_{\Gamma,\phi} \omega'$, iff $\omega_\beta = \omega'_\beta$ and either $\omega_i = \gamma_i$ or $\omega'_i = \neg\gamma_i$ for each $i = 1, \ldots, m$.*

In other words, one outcome is weakly preferred to another if the two have the same residual and the former satisfies all goal propositions satisfied by the latter. These comparisons make sense only for identical residuals because the agent may have arbitrary preferences over this attribute—by definition, the part of the outcome not covered by goals.

The goal preference order $\succsim_{\Gamma,\phi}$ is a partial preorder, actually a sub-order of the complete preference relation $\succsim$. For exact framings ϕ, the partial order takes the mathematical form of a collection of separate lattices, one for each distinct residual outcome ω_β. In this case, goal preference completely characterizes the preferences derivable from goals alone. We say that a preference order is *congruent* with a goal set Γ in ϕ if $goal(\gamma, \phi)$ holds according to the order for every $\gamma \in \Gamma$.

Proposition 3 *If ϕ is exact, then $\omega \succsim_{\Gamma,\phi} \omega'$ holds just in case $\omega \succsim' \omega'$ holds for every preference order $\succsim'$ congruent with the goal set Γ in ϕ.*

For inexact framings, goal preference is equivalent to agreement with every congruent preference order over an exact completion of the outcome set.

The incompleteness of the order $\succsim_{\Gamma,\phi}$ means that goals do not, by themselves, prescribe a unique choice of action in all circumstances. If one seeks to ensure unique rational choices, one must augment the goals with more detailed specifications of objectives—by specifying the relative strength of preference for competing goals and the form of the interactions underlying preference for combinations of goals. Nevertheless, the partial preference order induces a partial ranking of plans. In deterministic planning, we can extend goal preference directly to plans by defining $\pi \succsim_{\Gamma,\phi} \pi'$ iff $\rho(\pi) \succsim_{\Gamma,\phi} \rho(\pi')$. Expanding the definition, we see that $\pi \succsim_{\Gamma,\phi} \pi'$ means that π achieves all the goals that π' achieves and both produce the same residual outcome. Defining goal preferences in planning under uncertainty involves the notion of *stochastic dominance* (Fishburn and Vickson, 1978). See (Wellman *et al.*, 1991) for a detailed development in terms of utility theory.

Goal operations in planning

We seek, in the long term, to use our semantics for goals to provide a set of principles for designing and analyzing planning systems. For example, one may investigate whether the computational operations commonly applied to goal expressions, such as introducing and eliminating conjunctions and disjunctions of goals, only produce new goals from existing ones. If not, the choices made by planning agents may be incoherent with respect to the underlying preferences. In fact, our semantics reveals that these operations are not always valid.

Proposition 4 *Let $\phi = \langle \alpha, \alpha', \beta \rangle$ be a framing in which attributes α and α' represent the propositions γ and γ'. If we denote combinations of these attributes by boolean combinations of the attribute designators, then*

1. *$goal(\gamma, \phi)$ implies* *$goal(\gamma \wedge \gamma', \langle \alpha \wedge \alpha', \neg\alpha \wedge \neg\alpha', \alpha', \beta \rangle)$, and*
2. *$goal(\gamma, \phi)$ implies* *$goal(\gamma \vee \gamma', \langle \alpha \vee \alpha', \neg\alpha \vee \neg\alpha', \alpha', \beta \rangle)$,*

but the converse implications are invalid.

Invalidity of the converses indicates that subgoaling on conjunctions and disjunctions in AND/OR search need not always produce *bona fide* goals. Viewed semantically, the subgoals may have undesirable properties (side-effects) in addition to their relation to the compound goal. In general, preferences over composite propositions tell us little about preferences over their constituent parts.

Moreover, goalhood of a proposition implies goalhood of a conjunction or disjunction only for particular framings. In fact, even if we have $goal(\gamma, \phi)$ and $goal(\gamma', \phi)$ in an exact framing ϕ, it remains possible that $\gamma \wedge \gamma'$ not be a goal in some framing where $\alpha \wedge \alpha'$ is an attribute.

Although one cannot usually justify the behavior of a planning system using mere goals—due to the extreme incompleteness of the goal-preference order $\succsim_{\Gamma,\phi}$—the behavior can sometimes be validated conditional on additional restrictions, such as assumptions of preferential independence given combinations of propositions. We present a more detailed discussion of this approach elsewhere, along with analyses of other common goal manipulations (Wellman *et al.*, 1991).

Related work

Simon's (1955) initial critique of decision-theoretic rationality objected to straightforward descriptive uses of decision theory and to normative uses that fail to account for procedural factors. His theory of "satisficing" views goals as threshold "aspiration levels" that signal "satisfactory" (as opposed to optimal) levels of utility. Simon gives examples of how an agent might set acceptance levels given more precise description of preferences. The semantics presented here provides conditions that this mapping must satisfy, and in addition addresses the inverse problem: given goals, derive what one can about preferences. Our approach is to accept and exploit utility theory as the fundamental semantics for objectives, then consider bounded rationality in the design of decision-making procedures. The literature on satisficing does not appear to recognize this role for utility theory, even though it provides a way of relating procedural and substantive rationality and directly serves Newell's (1982) objective of knowledge level analysis.

Numerous authors have advocated and proposed techniques incorporating decision-theoretic ideas in AI planning, with Feldman and Sproull's (1977) work being perhaps the earliest and best known. For the most part these authors have either rejected goal-based specification of objectives entirely in favor of numeric utility functions, or have adopted *ad hoc* interpretations of goals, for example assigning them constant utility increments. While such interpretations are consistent with respect to our semantics, they also entail restrictive regularities in preference that we believe go far beyond the ordinarily intended preferential content of goals.

We attribute the paucity of previous work relating goals and preferences to the prevailing attitudes that either goals suffice for effective planning or that they represent trivial preferences (*i.e.,* the binary utility interpretation). Some work in AI, however, has attempted to combine notions of goals and utilities (Dean and Wellman, 1989; Farquhar, 1987; Haddawy and Hanks, 1990; Loui, 1990). In particular, Haddawy and Hanks (1990) present some methods for mapping between the two concepts in the context of planning under uncertainty. One major difference between their treatment and ours lies in the *ceteris paribus* condition in our definition of goalhood. While they also recognize the inappropriateness of preferring all outcomes satisfying the goal to all others, their approach deals with the problem by placing bounds on the utility difference among outcomes within each part of the partition. However, for problems with multiple goals or competing objectives, variations in other salient features of outcomes can defeat any fixed bounds on utility differences for a particular goal proposition.

Finally, we note that the semantics developed here formalize the methods in our previous work on decision-theoretic planning, which defined preference for a proposition by specifying a *positive qualitative influence* on utility (Wellman, 1990a). The use of qualitative influences in that work suggests how to extend our framework to account for preferences over ordinally scaled quantities in addition to propositions.

Conclusions

We have shown both how to give goals a nontrivial semantics in terms of decision-theoretic preferences, and how to construct preferences corresponding to specific sets of goals. The incompleteness of the preferences induced by goals formally establishes the inadequacy of goals as the sole basis for rational action. Despite their limitations, however, goals offer significant heuristic advantages over the utility functions developed in decision theory. The latter offer a more encompassing basis for rational action, but seem to require onerous computational expense, at least in straightforward mechanizations of decision-theoretic principles. The heuristic advantages of goals stem from the way planners use them to encode both preferential and intentional information. By fixing attributes of the outcome space, the intentional import of goals reduces the dimensionality of the utility function, focuses and organizes the search process, and provides a convenient skeleton for specifying control strategies. The preferences induced by goals, in turn, present a simpler decision problem than full expected utility maximization, by setting bounds on the search and by making the stopping criterion locally testable. Our preferential semantics provides an avenue for exploiting these heuristic advantages of goal-based planning representations without necessarily sacrificing decision-theoretic accountability.

The definition of goals in terms of preferences formalizes the intuition that goals are propositions that are preferred to their opposites, other things being equal. We demonstrated how for some goals this desirability depends on how one describes outcomes, and offered some suggestions for avoiding this sensitivity to representation. We also showed that while this semantics displays some intuitive properties, it also reveals that other seemingly natural planning operations are not always valid. To justify their systems' behavior, therefore, designers of planning architectures need either to provide further constraints on the meaning of goals or to find other means for expressing preference information. This highlights the importance of developing more refined languages for specifying the objectives of planning agents.

The semantics presented here constitutes part of a comprehensive decision-theoretic account of planning (see also (Wellman, 1990a; Wellman, 1990b)), and a more thorough treatment of the issue of goals and utilities is in preparation (Wellman *et al.*, 1991). We expect that much might be learned by developing planning architectures which combine goals with other preferences in a manner faithful to our semantics.

Acknowledgments We thank Jack Breese, Peter Haddawy, Steve Hanks, Ronald Loui, and Mark Peot for valuable discussions, and Tom Dean for useful suggestions on earlier drafts.

References

Cohen, Philip R. and Levesque, Hector J. 1990. Intention is choice with commitment. *Artificial Intelligence* 42:213–261.

Dean, Thomas and Wellman, Michael P. 1989. On the value of goals. In *Proceedings from the Rochester Planning Workshop*. Published as University of Rochester TR 284.

Doyle, Jon 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, Jon 1990. Rationality and its roles in reasoning (extended abstract). In *Proceedings of the Eighth National Conference on Artificial Intelligence*, Menlo Park, CA. AAAI Press. 1093–1100.

Farquhar, Peter H. 1987. Applications of utility theory in artificial intelligence research. In Sawaragi, Y.; Inoue, K.; and Nakayama, H., editors, *Toward Interactive and Intelligent Decision Support Systems, Volume 2*, Volume 286 of *Lecture notes in economics and mathematical systems*. Springer-Verlag. 155–161.

Feldman, Jerome A. and Sproull, Robert F. 1977. Decision Theory and Artificial Intelligence II: The hungry monkey. *Cognitive Science* 1:158–192.

Fishburn, Peter C. and Vickson, Raymond G. 1978. Theoretical foundations of stochastic dominance. In Whitmore, G. A. and Findlay, M. C., editors, *Stochastic Dominance: An Approach to Decision Making Under Risk*. D. C. Heath and Company, Lexington, MA.

Good, I. J. 1962. A five-year plan for automatic chess. In Dale, E. and Michie, D., editors, *Machine Intelligence 2*. Oliver and Boyd, London. 89–118.

Gorman, W. M. 1968. The structure of utility functions. *Review of Economic Studies* 35:367–390.

Haddawy, Peter and Hanks, Steve 1990. Issues in decision-theoretic planning: Symbolic goals and numeric utilities. In *Proceedings of the DARPA Workshop on Innovative Approaches to Planning, Scheduling, and Control*. 48–58.

Keeney, Ralph L. 1981. Analysis of preference dependencies among objectives. *Operations Research* 29:1105–1120.

Keeney, Ralph L. and Raiffa, Howard 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley and Sons, New York.

Loui, Ronald 1990. Defeasible specification of utilities. In Kyburg, Henry E. Jr.; Loui, Ronald P.; and Carlson, Greg N., editors, *Knowledge Representation and Defeasible Reasoning*. Kluwer Academic Publishers. 345–359.

Newell, Allen 1982. The knowledge level. *Artificial Intelligence* 18(1):87–127.

Savage, Leonard J. 1972. *The Foundations of Statistics*. Dover Publications, New York, second edition.

Simon, Herbert A. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69:99–118.

Wellman, Michael P. 1990a. *Formulation of Tradeoffs in Planning Under Uncertainty*. Pitman and Morgan Kaufmann.

Wellman, Michael P. 1990b. The STRIPS assumption for planning under uncertainty. In *Proceedings of the National Conference on Artificial Intelligence*, Boston, MA. American Association for Artificial Intelligence.

Wellman, Michael P.; Doyle, Jon; and Dean, Thomas 1991. Goals, preferences, and utilities: A reconciliation. In preparation.