In *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR-94)*, Bonn, Germany, May 24 - 27, 1994, pp 530 - 539.

Technical Report
R-205-II
Jan, 1994

Specification and Evaluation of Preferences under Uncertainty

Sek-Wah Tan and **Judea Pearl**
< *tan@cs.ucla.edu* > < *judea@cs.ucla.edu* >
Cognitive Systems Lab, Computer Science Department
University of California, Los Angeles, CA 90024
United States of America

Abstract

This paper describes a framework for specifying preferences in terms of conditional desires of the form "α is desirable if β", to be interpreted as "α is preferred to $\neg\alpha$ other things being equal in any β world". We demonstrate how such preference sentences may be interpreted as constraints on admissible preference rankings of worlds and how they, together with normality defaults, allow a reasoning agent to evaluate queries of the form "would you prefer σ_1 over σ_2 given ϕ" where σ_1 and σ_2 are action sequences. We also prove that by extending the syntax to allow for importance-rating of preference sentences, we obtain a language that is powerful enough to represent all possible preferences among worlds.

1 Introduction

This paper describes a framework for specifying planning goals in terms of preference sentences of the form "prefer α to $\neg\alpha$ if γ". Consider an agent deciding if she should carry an umbrella, given that it is cloudy. Naturally, she will have to consider the prospect of getting wet $\neg d$ (not dry), the possibility of rain r, that it is cloudy c, and so on. Some of the beliefs and knowledge that will influence her decision may be expressed in conditional sentences such as: "if I have the umbrella then I will be dry", $u \rightarrow d$, "if it rains and I do not have the umbrella then I will be wet", $r \wedge \neg u \rightarrow \neg d$ and "typically if it is cloudy, it will rain", $c \rightarrow r$. She may also have preferences like "I prefer to be dry", $d \succ \neg d$ and "I prefer not to carry an umbrella", $\neg u \succ u$. From the beliefs and preferences above, we should be able to infer whether to carry an umbrella if she observes that it is cloudy, assuming that keeping dry is more important to her than not carrying an umbrella.

The research reported in this paper concerns such decisions. Our aim is to eventually equip an intelligent

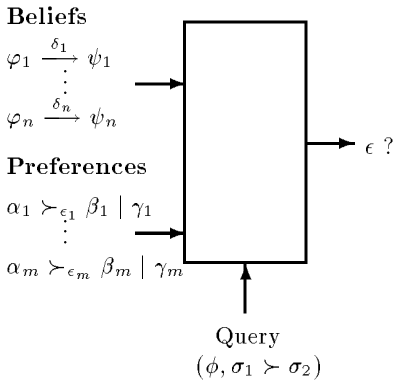

Figure 1: Schematic of the proposed system

autonomous artificial agent with decision making capabilities, based on two types of inputs: beliefs and preferences. Beliefs, some of which may be defeasible, will be specified by normality defaults like "if you run across the freeway then you are likely to die", written $run \rightarrow die$. Preferences may be encoded in conditional sentences such as "if it is morning then I prefer coffee to tea", written $coffee \succ tea \mid morning$. Figure 1 shows a schematic of the program. Each normality default $\varphi_i \xrightarrow{\delta_i} \psi_i$ and preference sentence $\alpha_i \succ_{\epsilon_i} \beta_i \mid \gamma_i$ will be quantified by an integer δ_i or ϵ_i which indicates the *degree* of the corresponding belief or preference. A larger degree implies a stronger belief or preference. The program will also accept queries in the form of $(\phi, \sigma_1 \succ \sigma_2)$, which stands for "would you prefer σ_1 over σ_2 given ϕ?". The output of the program is the degree ϵ to which the preference $\sigma_1 \succ \sigma_2$ holds in the context ϕ.

We take Bayesian decision theory and maximum expected utility [von Neumann and Morgenstern, 1947, Pearl, 1988, Keeney and Raiffa, 1976] as ideal norms for decision making. The problems with the theory are that it requires complete specifications of a probability distribution and a utility function and that the specifications are numeric. The problems with the complete specification of numeric probabilities had been considered and partly resolved in [Goldszmidt, 1992, Goldszmidt and Pearl, 1992]. The approach is to move from numeric probabilities to qualitative, order-of-magnitude abstractions and to use conditional statements of the form $\varphi \xrightarrow{\delta} \psi$ as a specification language that constrains qualitative probabilities. These constraints translate to a unique belief ranking $\kappa(\omega)$ on worlds that permits the reasoning agent to economically maintain and update a set of deductively closed beliefs. Pearl in [Pearl, 1993] addressed the problem of numeric utilities. Paralleling the order-of-magnitude abstraction of probabilities, he introduced an integer-valued utility ranking $\mu(\omega)$ on worlds that, combined with the belief ranking $\kappa(\omega)$, scores qualitative preferences of actions and their consequences. However, the requirement for the complete specification of the utility ranking remains problematic.

Here we propose a specification language which accepts conditional preferences of the form "if β then α is preferred to $\neg\alpha$", $\alpha \succ \neg\alpha \mid \beta$. A conditional preference of this form will also be referred to as a *conditional desire*, written $D(\alpha|\beta)$, which represents the sentence "if β then α is desirable". The output is the evaluation of a preference query of the form $(\phi, \sigma_1 \succ \sigma_2)$ where ϕ is any general formula while σ_1 and σ_2 are action sequences. The intended meaning of such query is "is σ_1 preferred to σ_2 given ϕ"? Our program is as follows. Each conditional desire $D(\alpha|\beta)$ is given *ceteris paribum* (CP) semantics; "α is preferred to $\neg\alpha$ other things being equal in any β-world". A collection of such expressions imposes constraints over *admissible* preference rankings $\pi(\omega)$. From the set of admissible rankings we select a subset of the most *compact* rankings $\pi^+(\omega)$, each reflecting maximal indifference. At the same time we use the normality defaults to compute the set of *believable* worlds $\{\omega \mid \kappa(\omega) = 0\}$ that may result after the execution of σ_i given ϕ. One way of computing the beliefs prevailing after an action is through the use of causal networks, as described in [Pearl, 1993]. To compare sets of believable worlds we introduce a preference relation between sets of worlds, called preferential dominance, that is derived from a given preference ranking $\pi(\omega)$. To confirm the preference query $(\phi, \sigma_1 \succ \sigma_2)$, we compare the set of believable worlds[1] resulting from executing σ_1 given ϕ to those resulting from executing σ_2 given ϕ, and test if the former *preferentially dominates* the latter in all the most compact preference rankings. A set of worlds W preferentially dominates V if and only if:

1. W provides more and better possibilities,
2. W provides less possibilities but excludes poorer possibilities or
3. W provides better alternative possibilities

when compared with V.

So far we have described the *flat* version of our language, where a degree is not associated with each conditional desire sentence $D(\alpha|\beta)$. We will show that the flat language is not sufficient for specifying all preference rankings. In particular we exhibit a preference ranking that is not the most compact admissible ranking with respect to any set of conditional desires. Also, by not specifying the relative importance of conditional desires, the flat language does not allow us to decide among preferences resulting from conflicting goals. To alleviate these problems we allow conditional desires to be quantified by a integer indicating the degree or strength of the desire. We prove that this quantified language is expressive enough to represent all preference rankings.

In the next section, we describe the language and the semantics for conditional desires. In section 3, we introduce preferential dominance between sets and show how a preference query may be evaluated. Quantified conditional desires are introduced in section 4 together with the sufficiency theorem. Related work is compared in section 5 and we conclude with a summary of the contributions of this paper.

2 Preference Specification

2.1 The Context

In this section we consider conditional desires of the form $D(\alpha|\beta)$ where α and β are well-formed formulas obtained from a finite set of atomic propositions $X = \{X_1, X_2, \ldots, X_n\}$ with the usual truth functionals $\wedge, \vee$ and $\neg$. Consider the desire sentence "I prefer to be dry", $D(d)$. This sentence may mean that

1. "d is preferred to $\neg d$ regardless of other things", or that
2. "d is preferred to $\neg d$ other things being equal" or
3. some intermediate reading.

In this paper we take the *ceteris paribum* (CP) reading which is "d is preferred to $\neg d$ other things being equal". Similarly, the interpretation for a conditional desire $D(\alpha|\beta)$ is "α is preferred to $\neg\alpha$ other things being equal in any β-world".

The first interpretation is not very useful, as shown by von Wright in [von Wright, 1963], in that it does not allow for two or more unconditional preference statements to exist consistently together. For example, the desire to be rich, $D(r)$ and the desire to be healthy,

[1] In general, "surprising worlds" should be considered as well, in case they carry extremely positive or negative utilities (e.g. getting hit by a car). But, to simplify the exposition, we consider only believable worlds. A system combining both likelihood and utility considerations, reflecting a qualitative version of the expected utility criterion, is described in [Pearl, 1993].

$D(h)$ will quickly run into a conflict when considering the worlds $r\overline{h}$ and $\overline{r}h$. This is because the world $r\overline{h}$ is preferred to $\overline{r}h$ by virtue of $D(r)$ and $\overline{r}h$ is preferred to $r\overline{h}$ by virtue of $D(h)$. The CP interpretation becomes reasonable in the light of this. Now we are going to question the CP interpretation.

Our first task is to explicate the meaning of $D(\alpha|\beta)$ in terms of preference constraints on pairs of worlds. Given the statement $D(\alpha)$, the CP interpretation imposes constraints only between worlds that agree on propositions that are not part of α. However to explicate what it means to be "part of α" it is insufficient to examine α syntactically, a semantic definition is required. For example, if $\omega = X_1 \wedge X_2 \wedge (\bigwedge_3^n X_i)$, $\nu = \neg X_1 \wedge \neg X_2 \wedge (\bigwedge_3^n X_i)$ and $\alpha = X_1$ we will conclude that $\omega \succ \nu$ is not sanctioned by CP, but if we were to write alpha as $X_1 \wedge (X_2 \vee \neg X_2)$ one might conclude that the preference above holds, because X_2 appears to be part of α and every thing else seems to be equal. To explicate this notion we say that a wff α *depends on* a proposition X_i if all wffs that are logically equivalent to α contain the symbol X_i. The set of propositions that α depends on is represented by $S(\alpha)$. This set is referred to as the *support* of α, written *support*(α) in [Doyle *et al.*, 1991]. The set of propositions that α does not depend on is represented by $\overline{S}(\alpha) = X \setminus S(\alpha)$. To explicate the notion of "other things being equal in any β-world", we say that two worlds *agree* on a proposition if they assign the same truth value to the proposition. Two worlds *agree* on a set of propositions if they agree on all the propositions in the set. We say that ω and ν are *S-equivalent*, written $\omega \sim_S \nu$ if ω and ν agree on the set $S \subseteq X$. Given a conditional desire $D(\alpha|\beta)$ and a β-world, ω, the worlds that have "other things being equal" in ω are those that are $\overline{S}(\alpha)$-equivalent to ω. We call $D(\alpha|\omega)$ a *specific* conditional desire if ω is a wff of the form $\bigwedge_1^n x_i$, where $x_i = X_i$ or $\neg X_i$. (As a convention we will use the same symbol ω to refer to the unique model of the wff ω.)

Every specific conditional desire imposes constraints on some set of worlds; we call that set the context.

Definition 1 (Context) *Let $D(\alpha|\omega)$ be a specific conditional desire. The* **context** *of $D(\alpha|\omega)$, $C(\alpha,\omega)$ is defined as*

$$C(\alpha,\omega) = \{\nu \mid \nu \sim_{\overline{S}(\alpha)} \omega\}. \tag{1}$$

We write $C_\gamma(\alpha,\omega)$ for $\{\nu \models \gamma \mid \nu \in C(\alpha,\omega)\}$ where γ is a wff.

In the umbrella example the support of $u \vee d$ is, $S(u \vee d) = \{u, d\}$ and the context of the specific conditional desire $D(u \vee d|udcr)$ is $\{udcr, u\overline{d}cr, \overline{u}dcr, \overline{u}\overline{d}cr\}$, the set of worlds which agree with $\omega = udcr$ on all propositions except for u and d. The constraints imposed by $D(u \vee d|udcr)$ are shown in figure 2, where

$$\begin{tikzcd} udcr \arrow[dr] & \overline{u}dcr \arrow[d] \\ u\overline{d}cr \arrow[r] & \overline{u}\overline{d}cr \end{tikzcd}$$

Figure 2: Constraints imposed by $D(u \vee d|udcr)$

the existence of an arrow $\omega \rightarrow \nu$ represents a preference constraint between ω and ν. The meaning of the direction of the arrow will be explained later.

Going from specific conditional desires to conditional desires, a conditional desire $D(\alpha|\beta)$ is interpreted as a conjunction of specific conditional desires $D(\alpha|\omega)$ over all models ω of β, $\bigwedge_{\omega\models\beta} D(\alpha|\omega)$. We note that $D(\alpha|\beta)$ may impose constraints on worlds that do not satisfy the condition β which may sound paradoxical. The reason being that each world fixes only $\overline{S}(\alpha)$, the atomic propositions which are not in α; however not all worlds that are constrained by $D(\alpha|\beta)$ are models of β; $\nu \in C(\alpha,\omega) \not\Rightarrow \nu \models \beta$. This stands contrary to [Doyle *et al.*, 1991] where conditional desires were restricted to apply only to the models of β. Consider the sentence, "I desire the light to be ON if it is night and the light is OFF", $D(l|n \wedge \neg l)$. Clearly such a sentence compares *night*-worlds in which the light is ON to those in which the light is OFF. The former does not satisfy the condition $\beta = n \wedge \neg l$. Such a reasonable sentence would be deemed meaningless in a restricted interpretation such as [Doyle *et al.*, 1991]. β does not act as a filter for selecting worlds to which the desired constraints apply, instead it identifies worlds in which the desires are satisfied.

2.2 Admissible Rankings

A preference ranking π is an integer-valued function on the set of worlds Ω. The intended meaning of a ranking is that the world ω is no less preferred than the world ν if $\pi(\omega) \geq \pi(\nu)$. Given a non-empty set of worlds, W, we write $\pi_*(W)$ for $\min_{\omega\in W} \pi(\omega)$ and $\pi^*(W)$ for $\max_{\omega\in W} \pi(\omega)$. If W is empty then we adopt the convention that $\pi_*(W) = \infty$ and $\pi^*(W) = -\infty$. The constraints imposed by a specific conditional desire $D(\alpha|\omega)$ translates into constraints over admissible preference rankings. The constraints are that every α-world in the context $C(\alpha,\omega)$ has a higher rank (is preferred) than any $\neg\alpha$-world in the same context.

Definition 2 (Admissibility of rankings) *Let D be a set of conditional desires. A preference ranking π is* **admissible** *with respect to a conditional desire $D(\alpha|\beta)$ if for all $\omega \models \beta$, $\nu \in C_\alpha(\alpha,\omega)$ and $\nu' \in C_{\neg\alpha}(\alpha,\omega)$ implies*

$$\pi(\nu) > \pi(\nu'). \tag{2}$$

A preference ranking π is **admissible** *with respect to D if it is admissible with respect to all conditional desires in D.*

If there exist a ranking that is admissible with respect to a set of conditional desires, D then we say that D is *consistent.* A trivial example of an inconsistent set is $\{D(u), D(\neg u)\}$. Another example of an inconsistent set is $\{D(\alpha), D(\neg\alpha|\beta)\}$. The proof will be given later. Figure 2 shows the three constraints imposed by the conditional desire $D(u \vee d|udcr)$. An arrow $\omega \rightarrow \nu$ represents the constraint $\pi(\omega) > \pi(\nu)$.

The principle of CP, though simple and reasonable, is still insufficient for drawing some conclusions we would normally draw from conditional desire sentences. Consider the sentence $D(d)$ "I desire to remain dry" in the original umbrella story. If this were truly the only desire we have, we should prefer every situation in which we are dry to any in which we are wet. No other consideration can get into conflict with this ramification. This conclusion is not sanctioned in the semantics considered thus far. For example we would not be able to deduce that the situation in which we are dry with the umbrella is preferred to the situation in which we are wet without the umbrella. The reason is that $D(d)$ does not impose any constraints between worlds that do not agree on any of the other propositions u, c or r. Although we do not want to deduce constraints between u and $\neg u$ worlds from the sole expression of the desirability of d, we would still want to be able to deduce a preference for d-worlds over $\neg d$-worlds by default if it is consistent to do so. This discussion suggests that in normal discourse we enforce additional constraints which are implicit in our reasoning. One such constraint is the principle of *maximal indifference.*

2.3 The Principle of Maximal Indifference

In [Goldszmidt, 1992] a distinguished ranking, the κ^+ ranking, was selected from among the admissible belief rankings. The κ^+ belief ranking assumes that every situation is as normal as possible, reflecting the principle of maximal ignorance. In the case of preferences the principle that we want to adopt is the principle of maximal indifference. We want to assume that there is no preference between two worlds unless compelled to be so by preferences that are explicated by the reasoning agent. From the set of admissible preference rankings we want to select a distinguished ranking which best capture the essence of the principle of maximal indifference. This ranking, the π^+ preference ranking, will minimize the difference in the preference ranks.

Definition 3 (The π^+ ranking) *Let D be a set of consistent set of conditional desires and let Π be the set of admissible rankings relative to D. A π^+ ranking is an admissible ranking that is* **most compact**, *that*

Table 1: Two most compact rankings

Worlds	π_1	π_2
abc	$m+2$	$m+2$
$\overline{a}bc$	$m+1$	$m+1$
$a\overline{b}c$	$m+1$	$m+1$
$\overline{a}\overline{b}c$	m	m
$ab\overline{c}$	$m+3$	$m+2$
$\overline{a}b\overline{c}$	$m+2$	$m+1$
$a\overline{b}\overline{c}$	$m+1$	m
$\overline{a}\overline{b}\overline{c}$	m	$m-1$

is

$$\sum_{\omega,\nu\in\Omega} |\pi^+(\omega) - \pi^+(\nu)| \leq \sum_{\omega,\nu\in\Omega} |\pi(\omega) - \pi(\nu)| \qquad (3)$$

for all $\pi \in \Pi$.

The π^+ rankings reflects maximal indifference[2] in the reasoning agent. Consider the extreme case where the set of desires D is empty. Without compactness, all preference rankings are admissible and no conclusions can be drawn. However with compactness we will select the "unique" ranking that ranks all worlds the same. In this way we make definite conclusions about the reasoning agent's lack of preferences among worlds.

In the umbrella example, if we have the sole desire $D(d)$ then the π^+ rankings are

$$\pi^+(\omega) = \begin{cases} m+1 & \text{if } \omega \models d \text{ and} \\ m & \text{otherwise.} \end{cases} \qquad (4)$$

where m is an integer. These preference rankings allow us to conclude that all worlds that satisfy d are preferred over all worlds that do not.

Although the π^+ ranking is unique in the above umbrella example it is not so in general. Consider the set $D = \{D(a|c), D(b|c), D(a|\neg c), D(a \wedge b|\neg c), D(a \vee \neg b|\neg c)\}$. The first two conditional desires impose the constraints

$$\pi(abc) > \pi(a\overline{b}c) > \pi(\overline{a}\overline{b}c)$$
$$\pi(abc) > \pi(\overline{a}bc) > \pi(\overline{a}\overline{b}c)$$

and the last three conditional desires dictate

$$\pi(ab\overline{c}) > \pi(a\overline{b}\overline{c}) > \pi(\overline{a}\overline{b}\overline{c}) > \pi(\overline{a}b\overline{c}).$$

Table 1 shows two admissible preference rankings of D. The sum of difference in ranks for both π_1 ad π_2 is 68 and that is the minimum sum achievable subject to the constraints. Therefore both π_1 ad π_2 are π^+ preference rankings of D.

[2] An alternative interpretation of maximal indifference can be developed whereby the distance $\pi(\omega) - \pi(\nu)$ cannot be reduced without either violating admissibility or increasing the difference between some other pair of worlds.

This is a simple and small example and the π^+ ranking can be easily computed. In the general case the conditional desires introduce a set of linear constraints between worlds of the form $\pi(\omega) - \pi(\nu) > 0$. The problem of finding the most compact preference ranking can be modeled as a nonlinear programming programming problem; minimizing

$$\sum_{\omega,\nu\in\Omega} |\pi(\omega) - \pi(\nu)|$$

subject to linear constraints of the form

$$\pi(\omega) - \pi(\nu) > 0.$$

There is no known efficient algorithm for solving the general nonlinear programming problem. However it is quite possible that this optimization problem is tractable for a restricted sublanguage of conditional preferences.

3 Evaluation of Preferences

3.1 The Role of Normality Defaults

So far we have paid no attention to normality defaults and this might lead us to counterintuitive behavior. Consider the preference query, "given that it is cloudy and raining, would you prefer to have an umbrella", $(cr, u \succ \neg u)$? If we have the sole desire $D(d)$ then we will certainly want to confirm the query despite the unlikely possibilities of remaining dry without the umbrella or being wet with the umbrella. Unless the knowledge base categorically excludes such scenarios as impossible, the semantics thus far will prevent us from the commonsensical conclusion to carry an umbrella. The purpose of normality defaults in the knowledge base is to identify such scenarios as unlikely. What we need is a system that on the one hand will keep esoteric situations as possibilities (just in case they become a reality) and on the other hand not let them interfere with mundane decision making. To disregard the unlikely scenarios, we compute the "believability" or likelihood of the worlds after the execution of actions, σ_i (u and $\neg u$ in this example) given some context, ϕ (cr in this example) and focus only on the worlds that are believable. An example of such a belief model is described in [Pearl, 1993][3]. We will assume that the output of this model is a belief ranking κ on worlds. We will write $\kappa(\phi; \sigma_i)$ to represent the ranking that results after the executing σ_i given context, ϕ. $\kappa^0(\phi; \sigma_i)$ will represent the set of believable worlds, namely the set of worlds for which $\kappa(\phi; \sigma_i) = 0$.

3.2 From Preferences on Worlds to Preferences on Sets

In a framework that tolerates imprecision and uncertainty, the consequence of the execution of an action may not be a specific world but a set of believable worlds. Thus to confirm a preference query we will need to define a preference relation between sets of worlds for example worlds in which we have an umbrella and worlds in which we do not have an umbrella. The straightforward approach would be to say that a set W (of believable worlds) is preferred over another set V if every world in W is preferred over any world in V. This criterion however is too restrictive. Consider the case where we have worlds u, v and w with ranks 0, 1 and 2 respectively. Let $W = \{u, v, w\}$ and $V = \{u, v\}$. In this example, the common possibilities u and v ensure that there is at least a world (u) in W that is not strictly preferred to a world (v) in V and vice versa. Therefore we are unable to determine any preference between the two sets because of the common possibility. However W offers all the possibilities that are available in V and in addition provides an additional possibility that is "better" than what is currently available in V. Intuitively we ought to prefer W to V.

Another consideration in determining the preferences between sets of worlds is the likelihoods of the worlds. This is the theme in Bayesian decision theory where the expected utilities, the sum of the utilities weighted by their corresponding probabilities, are compared and the set with the largest expected utility is preferred. Unfortunately the basic assumption of this paper was that the numeric probabilities and utilities are not available; what we have are order-of-magnitude approximations of probabilities and utilities which are expressed as normality defaults and conditional desires. Pearl in [Pearl, 1993] proposed an order of magnitude of abstraction of the maximum expected utility criterion. There are two problems with the proposal. An assumption in the proposal is that the scale of the abstraction of preferences is the same as the scale of the abstraction of beliefs. While this assumption could conceivably be valid when the utility ranks are explicitly specified, it is not justifiable when beliefs and preferences are specified in terms of normality defaults and conditional desires. The other problem is that the conclusions of the system are not invariant under a lateral shift of worlds along the preference scale (a linear translation of the utility ranking). The utility rankings, π and $\pi + 1$ may admit different conclusions in the system. This is problematic in our framework because lateral shifts of admissible preference rankings are always admissible since conditional desires impose only interval constraints among worlds. In this paper we take into account the likelihoods of the worlds by comparing worlds only when they have the same belief ranks of 0. All worlds of the same degree are considered to be equally believable.

In summary, when determining the preference between two sets, we will assume that the worlds in both sets are equally believable and will consider separately three types of worlds characterizing the compared set:

[3] In [Pearl, 1993] the computation of the post-action beliefs requires the use of a causal model.

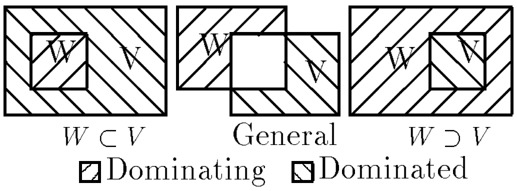

Figure 3: Interesting cases for $W \succ_\pi V$

the common possibilities, the additional possibilities and the excluded possibilities. Let us consider when we will prefer the set W over the set V (see figure 3) by imagining that the set V represents the possibilities that are currently available to us and the set W represents the set of new possibilities. Let us consider the case when $W \subset V$. Since W excludes some possibilities from V we have to compare these excluded possibilities (in $V \setminus W$) with the new possibilities offered by W. If the excluded possibilities are ranked lower than those that remain then W protects us from those excluded possibilities and we should prefer W to V. In the case when $V \subset W$, W provides more possibilities. If these additional possibilities (in $W \setminus V$) are ranked higher than the current possibilities, W provides an opportunity for improvement over the situation in V and again we should prefer W to V. In the general case, if W and V have some possibilities in common, then these common possibilities (in $W \cap V$) can be disregarded from consideration. If the additional possibilities (in $W \setminus V$) are ranked higher than the excluded possibilities (in $V \setminus W$) then we will prefer W to V. This motivates the definition of *preferential dominance*, a preference criterion between sets that depends on whether one set includes or overlaps the other.

Definition 4 (Preferential Dominance) *Let W and V be two subsets of Ω and let π be a preference ranking. We say that W π-**dominates** V, written $W \succ_\pi V$, if and only if $W \neq V$ and*

1. *$\pi_*(W) > \pi^*(V \setminus W)$ when $W \subset V$ or*
2. *$\pi_*(W \setminus V) > \pi^*(V)$ when $W \supset V$ or*
3. *$\pi_*(W \setminus V) > \pi^*(V \setminus W)$ otherwise.*

In figure 3, W π-dominates V, (written $W \succ_\pi V$), if the worlds in the dominating set are preferred over the worlds in the dominated set. Consider the example where we have the worlds u, v and w with preference ranks 0, 1 and 2 respectively. Let $W = \{u, w\}$ and $V = \{u, v\}$. In determining the preference between W and V, the common possibility u is disregarded and $\pi_*(W \setminus V) = 2 > \pi^*(V \setminus W) = 1$. Therefore $W \succ_\pi V$.

Table 2: Rankings in the umbrella example

Worlds ω	Preference ranking $\pi^+(\omega)$	Belief ranking $\kappa(\omega)$
$udcr$	$m+1$	0
$\overline{u}dcr$	$m+1$	> 0
$u\overline{d}cr$	m	> 0
$\overline{u}\overline{d}cr$	m	0
$udc\overline{r}$	$m+1$	> 0
$\overline{u}dc\overline{r}$	$m+1$	> 0
$u\overline{d}c\overline{r}$	m	> 0
$\overline{u}\overline{d}c\overline{r}$	m	> 0

3.3 Preferential Entailment

Let us consider the preference query "would you prefer σ_1 over σ_2 given ϕ?". In evaluating this query, we condition our beliefs on the context ϕ and compute the rankings that result after executing σ_i. To confirm the preference query $(\phi, \sigma_1 \succ \sigma_2)$, we compare the set of believable worlds resulting from executing σ_1 given ϕ with those resulting from executing σ_2 given ϕ, and test if the former *preferentially dominates* the latter in all the most compact preference rankings.

Definition 5 (Preferential Entailment) *Let D be a set of conditional desires and κ be some belief ranking on Ω. ϕ **preferentially entails** $\sigma_1 \succ \sigma_2$ given $\langle D, \kappa \rangle$, written $\phi \mathrel{|\!\!\sim} (\sigma_1 \succ \sigma_2)$, if and only if*

$$\kappa^0(\phi; \sigma_1) \succ_{\pi^+} \kappa^0(\phi; \sigma_2)$$

for all π^+ rankings of D.

Example
Let us reconsider the umbrella story where we need to verify the preference query "would you prefer to have the umbrella given that it is cloudy", $(c; u \succ \neg u)$? We have four atomic propositions, u - have umbrella, d - dry, c - cloudy and r - rain. Let us assume that we have the normality defaults, $\Delta = \{u \rightarrow d, r \wedge \neg u \rightarrow \neg d, c \rightarrow r\}$ and one unconditional desire, $D = \{D(d)\}$. For this example we will adopt the belief model in [Goldszmidt, 1992, Pearl, 1993]. First we process the defaults set Δ to get the resulting belief rankings $\kappa(\omega)$. Next, table 2 lists the possible worlds, given that it is cloudy, and gives the belief ranking $\kappa(\omega)$ and the π^+ preference ranking, where m is some fixed integer. $\kappa^0(c; u) = \{udcr\}$ and has rank $m+1$ while $\kappa^0(c; \neg u) = \{\overline{u}\overline{d}cr\}$ with rank m. Therefore the preference query $(c; u \succ \neg u)$ is confirmed.

4 Quantified Conditional Desires

A typical reasoning agent may have many desires. She may desire to be alive, $D(a)$, desire to be dry, $D(d)$ and also desire not to carry an umbrella, $D(\neg u)$. These

Table 3: Preference Rankings, π_1 and π_2

Worlds, ω	$\pi_1(\omega)$	$\pi_2(\omega)$
ab	2	2
$\overline{a}b$	0	1
$a\overline{b}$	1	1
$\overline{a}\overline{b}$	0	0

desires are not perceived as being equally important; being alive is more important than being dry and being dry is probably more important than not carrying an umbrella. In the specification language described so far there is no mechanism for indicating the varying degrees of preference. Let us examine the importance of having such degrees.

Suppose, in the umbrella example, that we have the desire $D_1(\neg u)$ in addition to the desire $D_2(d)$. These desires are quantified by a number indicating the strength of the preference. The strength of the desire to be dry is 2 which is stronger than the strength of the desire not to have the umbrella. In this case we will still expect the reasoning agent to confirm the preference query $(c; u \succ \neg u)$ as before. However, in the flat system where there is no consideration for the strength of the preferences, the constraints imposed by the two desires would yield

$$\pi^+(\omega) = \begin{cases} m+1 & \text{if } \omega \models d \wedge \neg u \text{ and} \\ m-1 & \text{if } \omega \models \neg d \wedge u \text{ and} \\ m & \text{otherwise} \end{cases}$$

as the most compact ranking. Now $\kappa^0(c; u)$ has the single world $udcr$ while $\kappa^0(c; \neg u)$ has the single world $\overline{u}\overline{d}cr$, both of rank m. This means that we are unable to confirm the obvious fact that one should carry an umbrella on a cloudy day, $(c; u \succ \neg u)$.

The unquantified specification language is also not expressive enough to express all possible preferences. Consider the preference rankings, π_1 and π_2, shown in table 3. For any set of conditional desires, π_2 is admissible whenever π_1 is admissible because the language does not allow us to impose a constraint between $a\overline{b}$ and $\overline{a}b$. Furthermore π_2 is more compact than π_1 because $\sum |\pi_1(\omega) - \pi_1(\nu)| = 7 > \sum |\pi_2(\omega) - \pi_2(\nu)| = 6$. Therefore π_1 cannot be the π^+ ranking for any set of conditional desires. This means that if π_1 represents our preferences among worlds, there is no way we can express these preferences exactly in terms of conditional desires alone.

To alleviate these weaknesses we extend the syntax of the specification language by quantifying a conditional desire with an integer ϵ which indicates the strength of the desire. A *quantified* conditional desire is a preference expression of the form $D_\epsilon(\alpha|\beta)$, where ϵ is a integer, read: "Given β, α is preferred to $\neg\alpha$ by ϵ".

Definition 6 (Quantified Admissibility) *Let D be a set of quantified conditional desires. A preference ranking π is said to be* **admissible** *with respect to a quantified conditional desire $D_\epsilon(\alpha|\beta)$ if for all $\omega \models \beta$, $\nu \in C_\alpha(\alpha, \omega)$ and $\nu' \in C_{\neg\alpha}(\alpha, \omega)$ implies*

$$\pi(\nu) \geq \pi(\nu') + \epsilon. \tag{5}$$

A preference ranking is **admissible** *with respect to D if it is admissible with respect to all desires in D.*

An unquantified conditional desire is assumed to have a default degree of $\epsilon = 1$.

Example with multiple desires

Let us reconsider the umbrella example assuming that we have two desires $D_2(d)$ and $D_1(\neg u)$. The degrees of these desires indicate that the desire to remain dry is more important by an order of magnitude than the discomfort of carrying an umbrella. The most compact preference ranking in this case is

$$\pi^+(\omega) = \begin{cases} m+3 & \text{if } \omega \models d \wedge \neg u \text{ and} \\ m+2 & \text{if } \omega \models d \wedge u \text{ and} \\ m+1 & \text{if } \omega \models \neg d \wedge \neg u \text{ and} \\ m & \text{otherwise} \end{cases}$$

The believable worlds are $\kappa^0(c; u) = \{udcr\}$ with rank $m+2$ and $\kappa^0(c; \neg u) = \{\overline{u}\overline{d}cr\}$ with rank $m+1$. This confirms the obvious conclusion $(c; u \succ \neg u)$ (with degree 1) which remain unsettled in the flat system.

Now we want to show that the quantified language is powerful enough to express all possible preference ranking.

Definition 7 (Conditional Desires of π) *Given a preference ranking π, the conditional desires* **entailed** *by π is the set $D^\pi = \{D(\alpha|\beta) \mid \pi$ is admissible with respect to $D(\alpha|\beta)\}$.*

We note that if a preference ranking π is admissible with respect to D_1 and D_2 then π is admissible with respect to $D_1 \cup D_2$. This means that π is admissible with respect to D^π and D^π is the largest set that has π as an admissible preference ranking.

Theorem 1 (Uniqueness) *Let π and μ be preference rankings. If μ is admissible with respect to D^π then*

$$\mu = \pi + k$$

for some constant integer k.

In other words two preference rankings entail the same set of conditional desires if and only if one is a lateral shift of the other.

Corollary 1 (Sufficiency of the Language) *For all preference rankings, π, there exists a set of quantified conditional desires, Π, such that π is the most compact ranking admissible with respect to Π. In fact π is unique up to a linear translation.*

If our preferences among worlds are represented by a preference ranking, then the sufficiency corollary tells us that our preferences may be completely specified by a set of quantified conditional desires.

One significant point to note is that the proof of the sufficiency corollary makes use of conditional desires that have negative degrees. This is somewhat unfortunate as conditional desires with negative degrees are not particularly intuitive. Another way of augmenting the expressiveness of the specification language is to allow for conditional preferences of the form $\alpha \succ \beta \mid \gamma$, "if γ then α is preferred to β". This will not be considered here.

Another problem with the ceteris paribum semantics is that it does not handle specificity of conditional preferences very well. For example the conditional desires $\{D(\alpha|\beta), D(\neg\alpha|\beta')\}$ is inconsistent whenever $\beta \supset \beta'$.

Theorem 2 (Specificity) *If α, β and β' are wffs and $\beta \supset \beta'$ then $\{D(\alpha|\beta), D(\neg\alpha|\beta')\}$ is inconsistent.*

In normal discourse, we have no difficulty accommodating general expressions of preferences which are subsequently qualified in more specific scenarios. For example I desire to be alive, $D(a)$, yet I am willing to die for some noble cause, $D(\neg a|c)$. In such a situation $D(\neg a|c)$, having a more specific condition, overrides the former unconditional desire, $D(a)$. Such other desirable behavior is sanctioned by a more recent interpretation of conditional desires which further weakens the CP semantics [Tan, 1994].

5 Comparison with Related Work

Verification of the assertability of conditional ought statements of the form "you ought to do A if C" is considered in [Pearl, 1993]. The conditional ought statement is interpreted as "if you observe, believe or know C then the expected utility resulting from doing A is much higher than that resulting from not doing A". The treatment in [Pearl, 1993] assumed that a complete specification of a utility ranking on worlds is available and that the scale of the abstraction of preferences is the same as the scale of the abstraction of beliefs. Another problem is that the conclusions of the system is not invariant under a lateral shift of the utility ranking; for example utility rankings π_1 and π_2, where $\pi_2(\omega) = \pi_1(\omega) + 1$, may admit different conclusions; which endows special status to worlds toward which one is indifferent.

Goal expressions were given preference semantics in [Wellman and Doyle, 1991] while relative desires were considered in [Doyle *et al.*, 1991]. These accounts are similar to our semantics for unquantified unconditional desires. However their treatment of conditional preferences (called restricted relative desires) of the form "given γ, α is preferred over β" is problematic. In particular the semantics forces us to conclude that we must be indifferent[4] to the inevitable. This fatalistic view shows itself in a theorem: "you must be indifferent to α, given α". Thus if you discovered that your car has been stolen then you must be indifferent to it. While some may subscribe to such a fatalistic attitude, our semantics here is more optimistic.

In [Boutilier, 1993], expressions of conditional preferences of the form "$I(\alpha|\beta)$ - if β then ideally α", are given modal logic semantics in terms of a preference ordering on possible worlds. $I(\alpha|\beta)$ is interpreted as "in the most preferred worlds where β holds, α holds as well". This interpretation places constraints *only* on the most preferred β-worlds, allowing only β-worlds that also satisfy α to have the same "rank". This contrasts with our ceteris paribum semantics which places constraints between pairs of worlds. In discussing the reasoning from preference expressions to actual preferences (preference query in our paper) Boutilier [Boutilier, 1993] suggests that the techniques in default reasoning (for handling irrelevance in particular) could be similarly applied to preferential reasoning. For example he suggests that worlds could be assumed to be as preferred or as ideal as possible which parallels the assumption made in computing the κ^+ belief ranking [Goldszmidt, 1992], that worlds are as normal as possible. While it is intuitive to assume that worlds would gravitate towards normality because abnormality is a monopolar scale, it is not at all clear that worlds ought to be as preferred as possible since preference is a bipolar scale. In our proposal there is no preference for either end of the bipolar preference scale. The π^+ rankings actually compacts the worlds away from the extremes thus minimizing unjustified preferences. The difference can be seen in the example shown in table 1. The compactness criterion selects two distinguished compact preference rankings π_1 and π_2. If worlds are assumed to be as preferred as possible then π_1 would be the sole distinguished preference ranking. It remains to be seen if the I operator corresponds closely with the common linguistic use of the word "ideally".

In [Pinkas and Loui, 1992] consequence relations are classified according to their boldness (or cautiousness). We may also employ a bolder (or more cautious) entailment principle which would correspond to a risk seeking (or risk averse) disposition.

6 Conclusion

In this paper we describe a framework for specifying preferences in terms of conditional desires of the form "α is desirable if β", to be interpreted as "α is preferred to $\neg\alpha$ when all else is fixed in any β world". We demonstrate how such preference sentences may be in-

[4] You are indifferent to α if you desire both α and $\neg\alpha$.

terpreted as constraints on admissible preference rankings of worlds and how they, together with normality defaults, allow a reasoning agent to evaluate queries of the form "would you prefer σ_1 over σ_2 given ϕ" where σ_1 and σ_2 could be either action sequences or observational propositions. We also prove that by extending the syntax to allow for importance-rating of preference sentences, we obtain a language that is powerful enough to represent all possible preferences among worlds. This work is an extension of [Pearl, 1993] and [Doyle *et al.*, 1991].

A Proofs

Lemma 1 (Common Contexts) $\nu \in C(\alpha,\omega) \Rightarrow C(\alpha,\omega) = C(\alpha,\nu)$.

Lemma 2 (Extreme worlds) *Let π be a preference ranking and let μ be admissible with respect to D^π. For all contexts C,*

$$\pi(\omega) = \max_{\nu\in C}\pi(\nu) \Rightarrow \mu(\omega) = \max_{\nu\in C}\mu(\nu)$$

and

$$\pi(\omega) = \min_{\nu\in C}\pi(\nu) \Rightarrow \mu(\omega) = \min_{\nu\in C}\mu(\nu)$$

Proof: Let $\omega \in C$ and x_i be X_i if $\omega \models X_i$ and $\neg X_i$ otherwise. By lemma 1 we may assume that $C = C(\alpha,\omega)$ for some wff α. Consider $\beta = \bigwedge_{X_i\in S(\alpha)} x_i$. If $\pi(\omega) = \max_{\nu\in C}\pi(\nu)$ then $D_0(\beta|\omega) \in D^\pi$. This implies that $\mu(\omega) \geq \mu(\nu)$ for all $\nu \in C$. Therefore $\pi(\omega) = \max_{\nu\in C}\pi(\nu) \Rightarrow \mu(\omega) = \max_{\nu\in C}\mu(\nu)$. If $\pi(\omega) = \min_{\nu\in C}\pi(\nu)$ then $D_0(\neg\beta|\omega) \in D^\pi$. This implies that $\mu(\omega) \leq \mu(\nu)$ for all $\nu \in C$. So $\pi(\omega) = \min_{\nu\in C}\pi(\nu) \Rightarrow \mu(\omega) = \min_{\nu\in C}\mu(\nu)$. □

Corollary 2 (Extreme worlds) *Let π be a preference ranking and let μ be admissible with respect to D^π.*

$$\pi(\omega) = \max_{\nu\in\Omega}\pi(\nu) \Rightarrow \mu(\omega) = \max_{\nu\in\Omega}\mu(\nu)$$

and

$$\pi(\omega) = \min_{\nu\in\Omega}\pi(\nu) \Rightarrow \mu(\omega) = \min_{\nu\in\Omega}\mu(\nu)$$

Given a preference ranking, we write ω_* for a world that has the minimum rank and ω^* for a world that has maximum rank.

Lemma 3 (Larger Admissible Differences) *Let π be a preference ranking and let μ be admissible with respect to D^π. For all $\omega \in \Omega$,*

$$\mu(\omega) - \mu(\omega_*) \geq \pi(\omega) - \pi(\omega_*).$$

Proof: We will prove by induction on m, the number of variables ω and ω_* disagree on. In the base case, if $m = 0$ then $\omega = \omega_*$. Therefore the lemma holds trivially. Let us assume that the lemma holds for $m = 0,\ldots,k-1$. Without loss of generality, we may assume that ω and ω_* disagree on $Y = \{X_1,\ldots,X_k\}$ and that $\omega \models x_i$ for $i = 1,\ldots,m$. If $\pi(\omega) = \pi(\omega_*)$ then the theorem holds by corollary 2. Therefore we may assume that $\pi(\omega) - \pi(\omega_*) > 0$. We consider the context, $C = C(\bigwedge_1^k x_i|\omega)$.

If we can find a world $\nu \sim_{X\setminus X_i} \omega$, $\nu \models \neg x_i$ such that $\pi(\omega) \geq \pi(\nu)$ then let $d = D_{\pi(\omega)-\pi(\nu)}(x_i|\omega) \in D^\pi$ and we also have d implies $\mu(\omega) - \mu(\nu) \geq \pi(\omega) - \pi(\nu)$. Otherwise, let ν be such that $\pi(\nu) = \max_{\nu'\in C}\pi(\nu')$ and $d = D_{\pi(\omega)-\pi(\nu)}(\bigwedge_1^k x_i|\omega) \in D^\pi$. In this case, by lemma 2, we also have d implies $\mu(\omega)-\mu(\nu) \geq \pi(\omega) - \pi(\nu)$. Now clearly, in both cases, $\nu \neq \omega$. This implies, by the induction hypothesis, that $\mu(\nu)-\mu(\omega_*) \geq \pi(\nu) - \pi(\omega_*)$. By adding the two inequalities, we get the desired inequality $\mu(\omega)-\mu(\omega_*) \geq \pi(\omega)-\pi(\omega_*)$. □

Lemma 4 (Smaller Admissible Differences) *Let π be a preference ranking and let μ be admissible with respect to D^π. For all $\omega \in \Omega$,*

$$\mu(\omega) - \mu(\omega_*) \leq \pi(\omega) - \pi(\omega_*).$$

Proof: For all worlds ω, $D_{\pi(\omega_*)-\pi(\omega)}(\neg\omega) \in D^\pi$. This implies $\mu(\omega) - \mu(\omega_*) \leq \pi(\omega) - \pi(\omega_*)$. □

Theorem 1 (Uniqueness) Let π and μ be preference rankings. If μ is admissible with respect to D^π then

$$\mu = \pi + k$$

for some constant integer k.

Proof: Lemmas 3 and 4 imply that $\mu = \pi + \mu(\omega_*) - \pi(\omega_*)$. □

Corollary 1 (Sufficiency of the Language) For all preference rankings, π, there exists a set of quantified conditional desires, Π, such that π is the most compact ranking admissible with respect to Π. In fact π is unique up to a linear translation.

Proof: The proof follows when we set Π to be D^π. □

Theorem 2 (Specificity) If α, β and β' are wffs and $\beta \supset \beta'$ then $\{D(\alpha|\beta), D(\neg\alpha|\beta')\}$ is inconsistent.

Proof: (By contradiction) Let us assume that $\{D(\alpha|\beta), D(\neg\alpha|\beta')\}$ is consistent. Let π be an admissible preference ranking, the world ω be such that $\omega \models \beta$ (note that $\omega \models \beta'$ as well since $\beta \supset \beta'$) and $C = C(\neg\alpha,\omega) = C(\alpha,\omega)$. By $D(\alpha|\beta)$ we have $C_\alpha \succ_\pi C_{\neg\alpha}$ and by $D(\neg\alpha|\beta')$ we have $C_\alpha \succ_\pi C_{\neg\alpha}$. This is a contradiction. □

Acknowledgements

We would like to thank two anonymous reviewers for their constructive comments and suggestions. The first author is supported in part by a scholarship from the National Computer Board, Singapore. The research was partially supported by Air Force grant #AFOSR 90 0136, NSF grant #IRI-9200918, Northrop Micro grant #92-123, and Rockwell Micro grant #92-122.

References

[Boutilier, 1993] Craig Boutilier. A modal characterization of defeasible deontic conditionals and conditional goals. In *Working Notes of the AAAI Spring Symposium Series*, pages 30–39, Stanford, CA, March 1993.

[Doyle *et al.*, 1991] John Doyle, Yoav Shoham, and Michael P. Wellman. The logic of relative desires. In *Sixth International Symposium on Methodologies for Intelligent Systems*, Charlotte, North Carolina, October 1991.

[Goldszmidt and Pearl, 1992] Moisés Goldszmidt and Judea Pearl. Reasoning with qualitative probabilities can be tractable. In *Proceedings of the* 8^{th} *Conference on Uncertainty in AI*, pages 112–120, Stanford, 1992.

[Goldszmidt, 1992] Moisés Goldszmidt. *Qualitative Probabilities: A Normative Framework for Commonsense Reasoning.* PhD thesis, University of California Los Angeles, Cognitive Systems Lab., Los Angeles, October 1992. Available as Technical Report (R-190).

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

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, 1988.

[Pearl, 1993] Judea Pearl. From conditional oughts to qualitative decision theory. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, Washington DC, July 1993.

[Pinkas and Loui, 1992] Gadi Pinkas and Ronald P. Loui. Reasoning from inconsistency: A taxonomy of principles for resolving conflict. In *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, pages 709–719, Cambridge, MA, October 1992.

[Tan, 1994] Sek-Wah Tan. Qualitative decision theory. Technical Report 214, University of California, Los Angeles, 1994.

[von Neumann and Morgenstern, 1947] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behaviour.* Princeton University Press, second edition, 1947.

[von Wright, 1963] Georg Henrik von Wright. *The Logic of Preference.* Edinburgh, 1963.

[Wellman and Doyle, 1991] Michael P. Wellman and Jon Doyle. Preferential semantics for goals. In *Proceedings of the Ninth National Conference on AI*, pages 698–703, Anaheim, CA, 1991.