Specification and Evaluation of Preferences for Planning 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 behavior in terms of preference sentences of the form "prefer α to $\neg\alpha$ 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, together with normality defaults, may be used as constraints on admissible preference-ranking of worlds and how they 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.

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, whether 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". 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 the individual will prefer to carry an umbrella if she observes that it is cloudy, assuming that being dry is more important to her than not carrying an umbrella.

The research reported in this paper concerns such decision making processes. Our aim is to eventually equip an intelligent 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". Preferences may be encoded in conditional sentences such as "if it is morning then I prefer coffee to tea", written *coffee* $\succ$ *tea* | *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.

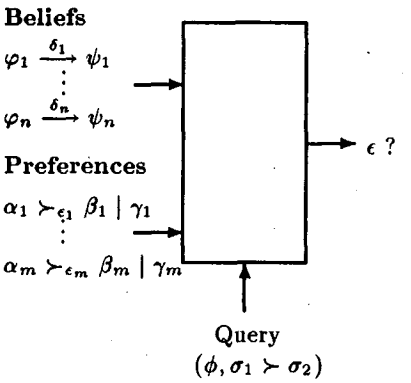

Figure 1: Schematic of the proposed system

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 state-

*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.

ments 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 may either be formulas or action sequences. The intended meaning of such query is "is σ_1 preferred to σ_2 given ϕ"? The idea is as follows. Each conditional desire $D(\alpha|\beta)$ is interpreted as "α is preferred to $\neg\alpha$ when all else is fixed 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 or observation of σ_i given ϕ. One way of computing the beliefs prevailing after an action or observation is through the use of causal networks, as described in [Pearl, 1993]. To compare the 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 or observing σ_1 given ϕ to those resulting from executing or observing 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

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

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 "d is preferred to $\neg d$ regardless of the truth of all the other propositions", or that "d is preferred to $\neg d$ given that all the other propositions are fixed" or some intermediate reading. In this paper we take the *ceteris paribum* reading which is "d is preferred to $\neg d$ given that all the other propositions are fixed". Similarly, the interpretation for a conditional desire $D(\alpha|\beta)$ is "α is preferred to $\neg\alpha$ when all else is fixed in any β-world".

To explicate the meaning of "all the other propositions", we need to distinguish between propositions that affect the desire from those that are irrelevant to the desire. We say that a wff α *depends on* a proposition X_i if all wffs that are logically equivalent to α contain X_i. The set of propositions that α depends on is represented by $I(\alpha)$. The set of propositions that α does not depend on is represented by $U(\alpha) = X \setminus I(\alpha)$. Accordingly, to explicate the notion of "all else is fixed 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 *U-equivalent*, written $\omega \sim_U \nu$ if ω and ν agree on the set $U \subseteq X$. Given a conditional desire $D(\alpha|\beta)$ and a β-world, ω, the worlds that have "all the other propositions fixed" by ω are those that are $U(\alpha)$-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.)

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

$$C(\alpha, \omega) = \{\nu \mid \nu \sim_{U(\alpha)} \omega\}. \quad (1)$$

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

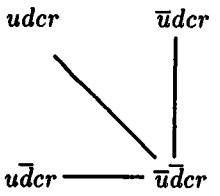

Figure 2: Constraints imposed by $D(u \vee d|udcr)$

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

As an example, $I(u \vee d) = \{u, d\}$ and the context of the specific conditional desire $D(u \vee d|udcr)$ is $\{udcr, ud\bar{c}r, \bar{u}dcr, \bar{u}d\bar{c}r\}$, 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 ω — ν represents a constraint between ω and ν.

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 $\pi^*(W) = -\infty$. A specific conditional desire $D(\alpha|\omega)$ imposes constraints on its context. The constraints are that every α-world in the context $C(\alpha,\omega)$ is preferred (has higher ranks) than any $\overline{\alpha}$-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, Π then we say that Π is *consistent*. An example of an inconsistent set is $D = \{D(u), D(\neg u)\}$. D is inconsistent because $D(u)$ implies that $\pi(u) > \pi(\neg u)$ while $D(\neg u)$ implies that $\pi(\neg u) > \pi(u)$.

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 is*

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

for all $\pi \in \Pi$.

The π^+ rankings reflects maximal indifference 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

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

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.

3 Preference Evaluation

Consider the preference query, "given that it is cloudy and raining, would you prefer to have an umbrella", $(cr, u \succ \neg u)$? Given that it is cloudy and raining, the possible scenarios for having the umbrella are $udcr$ and $ud\bar{c}r$ while the possible scenarios for not having the umbrella are $\bar{u}dcr$ and $\bar{u}d\bar{c}r$. If we have the desire $D(d)$, then the scenario $ud\bar{c}r$ where you get wet despite having the umbrella obviously has a preference ranking $\pi^+(ud\bar{c}r)$ which is strictly lower than $\pi^+(\bar{u}dcr)$, the rank of remaining dry without the umbrella. Although these scenarios are highly unlikely they are nevertheless possible and they prevent us from confirming the preference query, $(cr, u \succ \neg u)$. To disregard such unlikely scenarios, we expect our reasoning agent to consider normality defaults like "if I have the umbrella then I will be dry", $u \rightarrow d$ and "if it rains and I do not have the umbrella then I will be wet", $r \wedge \neg u \rightarrow \neg d$ and to compute the "believability" or likelihood of the worlds after the execution of actions or observation of evidence given some context, ϕ. An example of such a belief model is described in [Pearl, 1993][2]. 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 execution or observation of σ_i given context, ϕ. $\kappa^0(\phi;\sigma_i)$ will represent the set of believable worlds, namely the set of worlds for which $\kappa(\phi;\sigma_i)$ is minimal (least surprising).

In a framework that tolerates imprecision and uncertainty, the consequence of the execution of an action or the observation of evidence may not be a specific world but a set of believable worlds. Thus to confirm a preference query we will need to define preference between sets of worlds. 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 999 respectively. Let $W = \{u, w\}$ and $V = \{u, v\}$. The additional possibility w offered by W, which has rank 999 is preferred over the additional possibility v offered by V, which has a rank of 1. So in this case we would clearly prefer W over V as the additional possibility offered by W is so much better than the additional possibility offered by V. This preference

[2]The computation of the post-action beliefs in [Pearl, 1993] requires the use of a causal model.

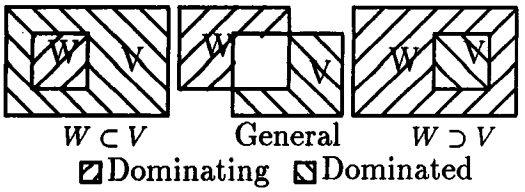

Figure 3: Interesting cases for $W \succ_\pi V$

however is not sanctioned by the criteria above. The reason for this is that the sets of: common possibilities, additional possibilities offered and excluded possibilities are not considered separately. This motivates the definition of *preferential dominance*, a preference criterion between sets that depends on whether one set includes or overlaps the other.

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. For example, if $W \subset V$ then we will still prefer W if the additional possibilities offered by V are worse than what we already have in W. On the other hand if $V \subset W$, we should prefer W over V if the possibilities excluded by V are all preferred to those that are left. In the general case, both V and W may offer some possibilities that are not offered by the other, in which case we simply compare these additional possibilities, while ignoring the common possibilities.

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.*

To confirm the preference query $(\phi, \sigma_1 \succ \sigma_2)$, we compare the set of believable worlds resulting from executing or observing σ_1 given ϕ to those resulting from executing or observing 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 $\phi \mid\!\sim (\sigma_1 \succ \sigma_2)$, 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 and Pearl, 1992, Pearl, 1993]. First we process the defaults set Δ to get the resulting belief rankings $\kappa(\omega)$. Next, table 1 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 $\kappa^0(c; \neg u) = \{\overline{u}dcr\}$ with rank m. Therefore the preference query $(c; u \succ \neg u)$ is confirmed.

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

Table 1: Rankings in the umbrella example

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 desires are not all 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 consequences of this limitation.

Suppose, in the umbrella example, that we have the desire $D(\neg u)$ in addition to the desire $D(d)$. These desires impose two sets of constraints on the preference rankings, yielding

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

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 $\overline{u}dcr$, both of rank m. This leads to counterintuitive indifference between $udcr$ and $\overline{u}dcr$, assuming that the desire to remain dry is more important that the desire not to carry an umbrella.

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 2. 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 preference among worlds then there is no way we can

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

Table 2: Preference Rankings, π_1 and π_2

express our 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*

$$\pi(\nu) \geq \pi(\nu') + \epsilon. \quad (4)$$

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 $\kappa^0(c;\neg u) = \{\overline{u}dcr\}$ with rank $m+1$. This confirms the preference query $(c; u \succ \neg u)$ (with degree 1).

Theorem 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.*

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. Another problem is that the conclusions of the system is not invariant under a translation 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.

In [Doyle *et al.*, 1991, Wellman and Doyle, 1991], preference semantics were given for goals and relative desires. The account is very similar to our semantics for unquantified unconditional desires. However the treatment of conditional preferences (called restricted relative desires) of the form "given γ, α is preferred over β" is problematic. In particular the semantics allows us to conclude that we must be indifferent[3] 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 (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 towards the middle of the scale. 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 have described a framework for specifying behavior in terms of preference sentences of the form "prefer α to $\neg\alpha$ if β", interpreted as "α is preferred to $\neg\alpha$ when all else is fixed in any β world". We demonstrate how such preference sentences, together with nor-

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

mality defaults, may be interpreted as constraints on the admissible preferences between worlds and how they allow the 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 the quantification of preference sentences, representing their importance, we have a language that is powerful enough to represent all possible preferences between worlds. This work is an extension of [Pearl, 1993] and [Doyle *et al.*, 1991].

A Proofs

Definition 7 *Given a preference ranking π, $D^\pi = \{d \mid \pi$ is admissible with respect to $d\}$.*

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 $\beta = \bigwedge_{X_i \in I(\alpha)} x_i$. 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 1 (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 1. 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 inequalitites, 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 $\mu(\omega) - \mu(\omega_*) \leq \pi(\omega) - \pi(\omega_*)$. □

Lemma 5 (Uniqueness) *Let π be a preference ranking. 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 $\mu = \pi + \mu(\omega_*) - \pi(\omega_*)$. □

Theorem 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: This theorem follows as a corollary of lemma 5 by setting Π to be D^π. □

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. Rank-based systems: A simple approach to belief revision, belief update and reasoning about evidence and actions. In *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, pages 661–672, Cambridge, MA, October 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.

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

[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.