Reprinted from *Proceedings of the Sixth International Symposium on Methodologies for Intelligent Systems (ISMIS'91)*, October, 1991, Z. W. Ras and M. Zemankova, editors, Berlin: Springer-Verlag, pp. 16–31.

A Logic of Relative Desire

(Preliminary Report)

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

Yoav Shoham
Department of Computer Science
Stanford University
Stanford, CA 94305
shoham@cs.stanford.edu

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

Abstract: Although many have proposed formal characterizations of belief structures as bases for rational action, the problem of characterizing rational desires has attracted little attention. AI relies heavily on goal conditions interpreted (apparently) as absolute expressions of desirability, but these cannot express varying degrees of goal satisfaction or preferences among alternative goals. Our previous work provided a relative interpretation of goals as qualitative statements about preferability, all else equal. We extend that treatment to the comparison of arbitrary propositions, and develop a propositional logic of relative desire suitable for formalizing properties of planning and problem-solving methods.

1 Introduction

Question your desires.—William Shakespeare

Standard theories of rational action take decisions of the agent to depend on beliefs about the relative desirability of the results of its available actions [2, 3]. The predominant approach to planning in artificial intelligence represents desires by conditions on the state of the world called *goals*. Each goal represents a partition of possible states into those satisfying and those not satisfying the goal. Goals serve a dual role in most planning systems, capturing aspects of intentions as well as desires [1]. Besides expressing the desirability of a state, adopting a goal typically represents some commitment to pursuing that state.

In perhaps the simplest interpretation of goals as desires, the states satisfying the goal are considered desirable in an *absolute* sense. That is, goal conditions define a partition of states into the desirable and the undesirable. However, this crude binary distinction proves inadequate in realistic situations. These satisfy objectives to varying degrees, and an absolute notion of desire cannot distinguish among alternative plans that ensure achievement of goals, nor among plans that fail to guarantee goal achievement.

We can support finer distinctions by expressing the *relative* desirability of alternate states. For example, we might hold that achieving conditions p and q would be more desirable than achieving neither, but that if we can achieve only one we would prefer to achieve p. Decision-theoretic preference orders express exactly these sorts of comparisons, and in related work [7], we provide a decision-theoretic semantics for goals in terms of preference orders and multiattribute outcome spaces. In our semantics, we relativize the notion of goal by restricting preference comparisons to fixed contexts. Specifically, we call a condition p a goal or desire if and only if (iff) p is preferred to $\neg p$, holding all other properties constant. Unlike the absolute interpretation, this relative concept admits multiple desires or goals. That is, desiring both p and q means preferring each one to its negation, holding the other constant.

In this paper, we extend our previous work in two primary directions. First, we consider relative desire between arbitrary conditions, in addition to that between a goal condition and its negation. Second, we formalize the concept of relative desire over conditions expressed in a propositional language. The resulting logic provides a general framework for representing and reasoning about desires. While such frameworks are common for beliefs, formal theories of desires have heretofore been conspicuously absent in artificial intelligence as well as philosophy [4].

We begin by introducing the basic notion of preference over models, which serves as the fundamental concept underlying our definition of relative desire over logical sentences. We follow by developing a collection of inference rules by which relative desires among some sentences entail desires over related sentences. We further extend the formalism by considering restrictions imposed by knowledge or constitution on preferences, and how relative desire varies with these restrictions.

2 Preference over models

Let $\mathcal{L}$ denote a logical language over a set of atoms $\mathcal{A}$ and the standard connectives $\neg$, $\wedge$, $\vee$, $\rightarrow$, and $\leftrightarrow$. The *literals* of $\mathcal{L}$ consist of just the atoms and their negations, that is, $\mathcal{A} \cup \{\neg a \mid a \in \mathcal{A}\}$. More generally, we write *literals*(A), for any subset $A \subseteq \mathcal{A}$, to mean $A \cup \{\neg a \mid a \in A\}$. In the following, we write $a, a', \ldots$ to denote atoms; $l, l', \ldots$ to denote literals; $c, c', d, \ldots$ to denote conjunctions of literals; and $p, q, r, \ldots$ to denote sentences in $\mathcal{L}$. We write *true* and *false* to denote arbitrary formulas of the form $p \vee \neg p$ and $p \wedge \neg p$, respectively.

A *model* of $\mathcal{L}$ assigns truth values to all atoms of $\mathcal{L}$, and so, by the usual logical rules, to all sentences in $\mathcal{L}$. We represent models by complete consistent sets of literals. Thus $m \subset \mathit{literals}(\mathcal{A})$ represents a model just in case exactly one of a and $\neg a$ is in m for each $a \in \mathcal{A}$, and we say that m makes a true (resp. false) just in case $a \in m$ (resp. $\neg a \in m$). (We may thus view a conjunction c of literals over different atoms as a sentence representing a partial model.) We write $\mathcal{M}$ to mean the set of all models of $\mathcal{L}$, and write $m, m', \ldots$ to denote individual models.

Following the standard definition, we say that a model m *satisfies* a sentence p, written $m \models p$, if the truth values it assigns to the atoms in p make p true; that is, $m \models a$ iff $a \in m$; $m \models p \wedge q$ iff $m \models p$ and $m \models q$; $m \models \neg p$ iff $m \not\models p$, etc. As usual, we write $p \models q$ to mean that every model satisfying p also satisfies q. If every model in $\mathcal{M}$ satisfies p, we write $\models p$ and say that p is *valid*.

A *proposition* is a set of models. We define $[\![p]\!]$, the proposition corresponding to p, by

$$[\![p]\!] \stackrel{\text{def}}{=} \{m \in \mathcal{M} \mid m \models p\}.$$

For a literal l, therefore, $[\![l \in \mathcal{M}]\!] = \{m \mid l \in m\}$.

Decision theory builds on the notion of preferences over outcomes or possible states of the world. It captures a notion of ideal rationality by requiring complete and consistent preferences, and by requiring that the agent always chooses actions leading to the most preferred outcomes.

In the logic of relative desire, we represent outcomes or states of the world by models, and so consider preferences over $\mathcal{M}$. We represent the agent's preferences by a preorder (a reflexive and transitive relation) $\succsim$ over $\mathcal{M}$, called the *preference order*. (We do not require here decision theory's standard assumption making $\succsim$ complete.) When $m \succsim m'$ we say that m is *weakly preferred* to m', which means that m is at least as desirable as m'. The strict preference order $\succ$ consists of the irreflexive part of $\succsim$, that is, $m \succ m'$ (m

is *preferred* to m') iff $m \succsim m'$ but $m' \not\succsim m$. When both $m \succsim m'$ and $m' \succsim m$, we say the two models are *indifferent*, and write $m \sim m'$.

Much work in decision theory concerns conditions under which one can represent $\succsim$ by order-preserving, real-valued *utility functions*, and with identifying regularities in preferences that justify utility functions with convenient structural properties [3]. Although we expect that utility theory will have much to offer for a calculus of desires for reasoning systems, our logic relies only on the ordinal preference relation, not on any numerical representations.

We may lift the preference order over models to the set of propositions $2^{\mathcal{M}}$ by defining $P \succsim Q$, for $P, Q \subseteq \mathcal{M}$, to hold iff $m \succsim m'$ for each $m \in P$ and $m' \in Q$, and by defining the lifted relations $\succ$ and $\sim$ similarly. Clearly, if $P \succ Q$, then P and Q must be disjoint. We may further translate the preference order into an order over sentences by saying that $p \succsim q$ holds iff $[\![p]\!] \succsim [\![q]\!]$.

The relation over propositions or sentences defined by simply lifting and translating the fundamental preference order, however, fails to support useful ways of combining multiple preferences, and hence cannot provide an adequate semantics for relative desires. For example, suppose we represent the desirability of p by the relation $p \succsim \neg p$. Then having several desirable sentences prevents us from distinguishing all but the extreme cases, as shown by the following result.

Theorem 1 (No tradeoffs) *Suppose $p_i \succsim \neg p_i$ for $0 \leq i \leq n$, with p_i and p_j logically independent for $i \neq j$. Then*

1. *Any model that satisfies every p_i is weakly preferred to any model that does not,*
2. *Any model that satisfies at least one p_i is weakly preferred to any model that falsifies every p_i, and*
3. *All models that satisfy some p_i and falsify some p_j are indifferent.*

Proof: By definition, the hypothesis entails $[\![p_i]\!] \succsim [\![\neg p_i]\!]$ for $0 \leq i \leq n$.

1. Let m satisfy all p_i and m' falsify sentence p_j. Therefore $m \in [\![p_j]\!]$ and $m' \in [\![\neg p_j]\!]$, so $m \succsim m'$.

2. Let m satisfy sentence p_j and m' falsify all p_i. Therefore $m \in [\![p_j]\!]$ and $m' \in [\![\neg p_j]\!]$, so $m \succsim m'$.

3. By pairwise logical independence, we may find a model $m_{i,j}$ satisfying sentence p_i and falsifying p_j for any $0 \leq i, j \leq n$, $i \neq j$. Because $m_{i,j} \in [\![p_i]\!]$ and $m_{k,i} \in [\![\neg p_i]\!]$, we

conclude $m_{i,j} \gtrsim m_{k,i}$ for all $k \neq i$. Similarly, $m_{k,i} \gtrsim m_{l,k}$ for all $l \neq k$. By transitivity of weak preference, we have $m_{i,j} \gtrsim m_{l,k}$. A symmetric argument yields $m_{l,k} \gtrsim m_{i,j}$, hence $m_{i,j} \sim m_{l,k}$. □

This result means that the simplistic translation of preference over models to desirability of sentences distinguishes only three degrees of multiple goal satisfaction: total success, partial success, and total failure. While one often wishes to consider tradeoffs among partial successes, this interpretation makes them preferentially indistinguishable.

3 Relative desire

To obtain a semantic account of desire that permits differentiation of partial satisfactions of multiple desires, we relativize the absolute interpretation to specific contexts and say that a sentence represents a desire if it is preferred to its contrary, other things equal. Understood in this way, worlds corresponding to a desire need not be preferred no matter what, only within each fixed context.

The *support* of a sentence p, denoted *support*(p), consists of the minimal set of atoms determining the truth value of p, for example, the atoms appearing in an irredundant sum-of-products sentence logically equivalent to p (that is, a sentence in which each product term is a prime implicant of p and in which removing any product term destroys the logical equivalence with p). We say that specific sentences have *disjoint support* if their respective supports do not overlap.

Definition 1 (Model equivalence) *We say that m and m' are* equivalent modulo p, *written $m \equiv m' \bmod p$, iff they are the same outside of support(p), or formally*

$$m \setminus (\mathit{literals}(\mathit{support}(p)) = m' \setminus (\mathit{literals}(\mathit{support}(p)).$$

Definition 2 (Model modification) *The set of* modifications of a model m making p true, *written $m[p]$, is the set of models of p which assign the same truth values as m to all atoms other than those on which p depends, or formally,*

$$m[p] \stackrel{\text{def}}{=} \{m' \in [\![p]\!] \mid m \equiv m' \bmod p\}.$$

Definition 3 (Relative desire) *We say that p is* desired over q, *written $p \trianglerighteq q$ and read briefly as "p over q," iff for each $m \in \mathcal{M}$, $m' \in m[p \wedge \neg q]$, and $m'' \in m[\neg p \wedge q]$ we have $m' \gtrsim m''$. We say that p is* strictly desired over q, *written $p \triangleright q$, iff $p \trianglerighteq q$ but not $q \trianglerighteq p$.*

That is, p is desired over q just in case any model making p true and q false is weakly preferred to any model making p false and q true, whenever the two models assign the same truth values to all atoms not relevant to those supporting p and q. Furthermore, p is strictly desired over q just in case this preference is strict for some pair of models.

Theorem 2 (Logical invariance) *If $\models p \leftrightarrow p'$, then $m[p] = m[p']$, and if $\models q \leftrightarrow q'$ as well, then $p \trianglerighteq q$ iff $p' \trianglerighteq q'$.*

We omit the obvious proof and a few other proofs due to space limitations.

Theorem 3 (General contraposition) *If $p \wedge r \trianglerighteq q \wedge r$, then $\neg q \wedge r \trianglerighteq \neg p \wedge r$.*

Theorem 4 (General reflexivity) *If $\models p \rightarrow p'$, then $p \trianglerighteq p'$ and $p' \trianglerighteq p$.*

This means that relative desire does not distinguish sentences from stronger or weaker conditions (including *true* and *false*), so the interesting cases of relative desire all concern relations among logically independent conditions.

We extend the model modification operation to sets of models $M \subseteq \mathcal{M}$ by defining

$$M[p] \stackrel{\text{def}}{=} \bigcup_{m \in M} m[p].$$

We may thus write serial modifications by left-associating the modification operator, with $m[p][q] = (m[p])[q]$.

Lemma 5 (Covering) *If $\mathit{support}(p) \subseteq \mathit{support}(q)$ and $m \equiv m' \bmod p$, then $m \equiv m' \bmod q$ and $m[q] = m'[q]$.*

Proof: Suppose $\mathit{support}(p) \subseteq \mathit{support}(q)$. Since

$$m \setminus \mathit{literals}(\mathit{support}(p)) = m' \setminus \mathit{literals}(\mathit{support}(p)),$$

it follows that

$$m \setminus \mathit{literals}(\mathit{support}(q)) = m' \setminus \mathit{literals}(\mathit{support}(q)),$$

hence $m[q] = m'[q]$. □

Lemma 6 (Reduction) *$\mathit{support}(p) \subseteq \mathit{support}(q)$ iff for all m, $m[p][q] = m[q]$.*

Lemma 7 (Conjunctive decomposition) *If p and p' have disjoint support, then*

$$m[p \wedge p'] = m[p][p'] = m[p'][p].$$

Proof: Expanding the definitions yields

$$\begin{aligned} m[p \wedge p'] &= \{m' \in [\![p]\!] \cap [\![p']\!] \mid m \equiv m' \bmod p \wedge p'\}, \text{ and} \\ m[p][p'] &= \{m'' \mid m' \in [\![p]\!] \wedge m'' \in [\![p']\!] \wedge m \equiv m' \bmod p \wedge m' \equiv m'' \bmod p'\}. \end{aligned}$$

Because p and p' have disjoint support, the conditions in the second expression entail $m'' \in [\![p]\!]$ and $m \equiv m'' \bmod p \wedge p'$, making the two expressions equivalent. A symmetric argument yields equality with $m[p'][p]$. □

One cannot drop the hypothesis of disjoint support since the equalities fail when $p = \neg p'$ or when $p = l \vee l'$ and $p' = l \vee l''$.

Lemma 8 (Disjunctive decomposition) *If p and p' have disjoint support, then*

$$m[p \vee p'] = m[p \wedge p'] \cup m[p \wedge \neg p'] \cup m[\neg p \wedge p'].$$

Proof: According to the definition,

$$m[p \vee p'] = \{m' \in [\![p]\!] \cup [\![p']\!] \mid m \equiv m' \bmod p \vee p'\}.$$

By the assumption of disjoint support and Lemma 5, model equivalence modulo $p \vee p'$ is identical to equivalence modulo $p \wedge p'$, as well as equivalence modulo $p \wedge \neg p'$ and $\neg p \wedge p'$. Noting that

$$[\![p]\!] \cup [\![p']\!] = ([\![p]\!] \cap [\![p']\!]) \cup ([\![p]\!] \cap [\![\neg p']\!]) \cup ([\![\neg p]\!] \cap [\![p']\!]),$$

we can decompose the expression for $m[p \vee p']$ into three conjunctive parts as required. □

One cannot drop the hypothesis of disjoint support since the equality fails when $p = \neg p'$.

Lemma 9 (Unique modification) *If p and c have disjoint support and $m \models c$, then $m[p \wedge c] = m[p]$.*

Proof: By Lemma 7, we have $m[p \wedge c] = m[p][c]$. Since $m'[c]$ consists of a singleton for every $m' \in \mathcal{M}$, and since $c \subseteq m$ for every $m \in [\![c]\!]$, we have $m' \in m[p]$ iff $m'[c] = \{m'\}$. □

Using these results, we can derive a canonical expression for any model modifications over a sentence expressed as a disjunction of conjunctions with disjoint support.

Theorem 10 (Amalgamation) *Suppose p and q have support disjoint from r. If $p \wedge r \trianglerighteq q \wedge r$ and $p \wedge \neg r \trianglerighteq q \wedge \neg r$, then $p \trianglerighteq q$.*

Proof: By Lemma 7 and disjoint support, the hypothesized relative desires imply, for all m, that $m[r][p \wedge \neg q] \gtrsim m[r][\neg p \wedge q]$ and $m[\neg r][p \wedge \neg q] \gtrsim m[\neg r][\neg p \wedge q]$. Then given m', choose m such that $m' \in m[r]$ (if $m' \in [\![r]\!]$) or $m' \in m[\neg r]$ (if $m' \in [\![\neg r]\!]$). In either case, $m'[p \wedge \neg q] \gtrsim m'[\neg p \wedge q]$, and so $p \trianglerighteq q$. □

Theorem 11 (Refinement) *If p, q, and c have disjoint support, then $p \wedge c \trianglerighteq q \wedge c$ whenever $p \trianglerighteq q$.*

Proof: The relation $p \wedge c \trianglerighteq q \wedge c$ holds if $m[p \wedge \neg q \wedge c] \gtrsim m[\neg p \wedge q \wedge c]$ for each m. By Lemma 7 and disjoint support, this condition reduces to $m[c][p \wedge \neg q] \gtrsim m[c][\neg p \wedge q]$. Because c is a conjunction of literals, Lemma 9 indicates the modification $m[c]$ is unique for any m; let that model be m'. The hypothesis $p \trianglerighteq q$ entails that for all m', $m'[p \wedge \neg q] \gtrsim m'[\neg p \wedge q]$, and hence the conclusion is established. □

Theorem 12 (Conditional transitivity) *Assuming that c, c', and p have disjoint support, if $c \trianglerighteq p$ and $p \trianglerighteq c'$, then $c \wedge p \trianglerighteq c' \wedge p$.*

Proof: By Theorem 11, $c \trianglerighteq p$ entails $c \wedge c' \trianglerighteq p \wedge c'$, and $p \trianglerighteq c'$ entails $p \wedge c \trianglerighteq c' \wedge c$. Thus for all m, $m[c \wedge c' \wedge \neg p] \gtrsim m[\neg c \wedge c' \wedge p]$ and $m[c \wedge \neg c' \wedge p] \gtrsim m[c \wedge c' \wedge \neg p]$, so by transitivity of $\gtrsim$, we have $m[c \wedge \neg c' \wedge p] \gtrsim m[\neg c \wedge c' \wedge p]$, exactly the condition needed for $c \wedge p \trianglerighteq c' \wedge p$. □

Theorem 13 (Literal transitivity) *Assuming that l, l', and p have disjoint support, if $l \trianglerighteq p$ and $p \trianglerighteq l'$, then $l \trianglerighteq l'$.*

Proof: Assume that $l \trianglerighteq p$ and $p \trianglerighteq l'$. By Theorem 12, $l \wedge p \trianglerighteq l' \wedge p$. By applying the same theorem to the contraposition of the hypothesized desires, we obtain $\neg l' \wedge \neg p \trianglerighteq \neg l \wedge \neg p$. By Theorem 3, this yields $l \wedge \neg p \trianglerighteq l' \wedge \neg p$. Amalgamating these two cases by Theorem 10 results in $l \trianglerighteq l'$. □

Note that transitivity does not hold in general. For example, by Theorem 4, we have $p \trianglerighteq$ *true* and *true* $\trianglerighteq q$ for any p and q. But if $\mathcal{A} = \{p, q\}$ and $\{\neg p, q\} \succ \{p, \neg q\}$, then $p \trianglerighteq q$ is false. (Assuming $p \trianglerighteq (p \vee r) \wedge q$ and $(p \vee r) \wedge q \trianglerighteq r$ provides a less trivial example.)

We obtain transitivity in Theorems 12 and 13 by restricting the form and interrelations of the sentences involved. The proofs of these results exploit such restrictions to ensure the uniqueness of modifications to a given model. We impose similar restrictions in application of the inference rules sanctioned by Theorem 11 and other results below.

In [8] we present an alternative formalization which requires that modifications of models minimize the changes made in the initial model. In the simplest case, this means using a model rather than its modifications when the model already makes the sentence true. In more complex cases, we want to pick out worlds in the modification that change as small a set of literals as possible (small in the set inclusion sense). More generally, we employ comparative similarity orders like those used in theories of counterfactuals and belief revision. Elsewhere, we relate the notion of relative desire to formalizations of conditional and deontic logics.

4 Propositional desires

In our semantics for goals viewed as desires [7], we represent states as vectors of attributes. We designate particular conditions as binary attributes, and define goalhood as preference for those conditions, holding all other attributes constant. We may reformulate that semantics in the logic of relative desire to obtain meanings for individual propositional desires, which we represent as sentences desired over their negations.

Definition 4 (Propositional desire) *We say that p is a desire, and write $desire(p)$, just in case $p \trianglerighteq \neg p$. We say that p is a strict desire, and write $DESIRE(p)$, just in case $p \vartriangleright \neg p$.*

Thus this definition picks out desires by comparing all of the possible ways of making sentences true and false while holding all else equal.

Corollary 14 (Logical invariance) *If $\models p \leftrightarrow q$, then $desire(p)$ iff $desire(q)$.*

Corollary 15 (Negation of desires) *If $desire(p)$, then not $DESIRE(\neg p)$.*

Corollary 16 (Trivial desires) *The extremal sentences $p \vee \neg p$ and $p \wedge \neg p$ are desires but not strict desires.*

It remains to be seen whether any reasonable semantics can be found which does not make desires of *true* and *false*.

Theorem 17 (Sets of desires) *Suppose* $\hat{p} = \{p_0, \ldots, p_n\}$ *and* $desire(p_i)$ *(or equivalently,* $p_i \trianglerighteq \neg p_i$*) for* $0 \leq i \leq n$*, where* p_i *and* p_j *have disjoint support for* $i \neq j$*. Then*

1. *Any model* m *is weakly preferred to any* m' *obtained by modifying* m *to falsify some desires (i.e., if* $\hat{p}' \subseteq \hat{p}$ *and* $m' \in m[\neg \bigvee \hat{p}']$*, then* $m \succsim m'$*), and*
2. *For any pair of models* m *and* m' *not related by such a modification (i.e., if* $\hat{p}' \subseteq \hat{p}$*, then* $m' \notin m[\neg \bigvee \hat{p}']$*), there is some preference order* $\succsim$ *compatible with the desires such that* $m \succ m'$*.*

Proof: 1. We proceed by induction on the size of $\hat{p}'$. In the base case, $|\hat{p}'| = 0$, so $m' = m$, and hence $m \succsim m'$ by the reflexive property of weak preference. For the inductive step, suppose the result holds for $|\hat{p}'| \leq k \leq n$. Choose a set $\hat{p}'$ of size k, and an element p^* from $\hat{p} \setminus \hat{p}'$. Let $\hat{p}'' = \hat{p}' \cup \{p^*\}$, and note that $m[\neg \bigvee \hat{p}''] = m[(\neg \bigvee \hat{p}') \wedge \neg p^*]$, which, by disjoint support and Lemma 7, is equivalent to $m[\neg \bigvee \hat{p}'][\neg p^*]$. Therefore, for any $m'' \in m[\neg \bigvee \hat{p}'']$, there is some $m' \in m[\neg \bigvee \hat{p}']$ such that $m'' \in m'[\neg p^*]$. Because $m' \in m'[p^*]$ (since p^* is satisfied in m and not falsified by $\hat{p}'$), $m'' \in m'[\neg p^*]$, and $m' \equiv m'' \bmod p^*$, *desire*(p^*) directly entails $m' \succsim m''$. By the inductive step we have $m \succsim m'$, and so by transitivity $m \succsim m''$.

2. We show that the first part of the theorem exhausts the requirements on $\succsim$. The conditions *desire*(p_i) require that $m' \succsim m''$ whenever $m' \in m[p_i]$, $m'' \in m[\neg p_i]$, and $m' \equiv m'' \bmod p_i$. But in this case, $m'' \in m'[\neg p_i]$, and so this requirement is covered by the modification relation. To account for the reflexive property of weak preference, note that $m \in m[\neg \bigvee \emptyset]$, and hence this is also covered by the condition. Finally, we consider transitivity. If $m'' \in m'[\neg \bigvee \hat{p}'']$ and $m' \in m[\neg \bigvee \hat{p}']$, then $m'' \in m[\neg \bigvee (\hat{p}' \cup \hat{p}'')]$. Thus, no new preferences are drawn from transitivity, and the consequences are exhausted. Therefore we are free to order arbitrarily any pair of models m, m' such that $m' \notin m[\neg \bigvee \hat{p}']$. □

Comparing this result with Theorem 1 highlights the difference between relative desire and absolute preference among model sets. Whereas absolute preference collapses partially satisfied models into a single indifference class, relative desire allows for finer distinctions. The weak constraints placed by desires on preference orders mean that desires 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 desires with more detailed specifications of relative desires among propositions.

Theorem 18 (Entailment of desires) *If* $desire(l)$ *and* $l' \trianglerighteq l$*, then* $desire(l')$*.*

PROOF: If $l = l'$ or $l = \neg l'$, the result holds trivially. Otherwise, l and l' have disjoint support. By Lemma 7, $m[l \wedge l'] = m[l'][l]$, and $m[\neg l \wedge l'] = m[l'][\neg l]$. Since $m[l']$ is unique, from *desire*(l) it follows that $m[l'][l] \succsim m[l'][\neg l]$. From $l' \trianglerighteq l$ (and Lemma 7), we have $m[l'][\neg l] \succsim m[\neg l'][l]$. A transitive combination of these preferences yields $m[l'][l] \succsim m[\neg l'][l]$, and rearranging, $m[l][l'] \succsim m[l][\neg l']$. A symmetric argument, exploiting the uniqueness of $m[\neg l']$, yields $m[\neg l][l'] \succsim m[\neg l][\neg l']$. Since for every m' there is some m such that $m' \in m[l]$ or $m' \in m[\neg l]$, the two cases entail *desire*(l'). □

One may use the logic of relative desire to investigate principles for designing and analyzing planning systems. For example, many planners produce new goals by introducing and eliminating conjunctions and disjunctions of existing goals. Given the logic of relative desire, we may ask whether such operations are sound with respect to our semantics, when goals are interpreted as desires. In fact, the logic reveals that these operations are not always valid. For example, if $\mathcal{A} = \{p, q\}$ and $\{p, q\} \succ \{\neg p, \neg q\} \succ \{p, \neg q\} \succ \{\neg p, q\}$, then *desire*$(p \wedge q)$ but neither *desire*(p) nor *desire*(q) holds. If the planner uses unsound operations on goals, then either it risks making choices during plan execution that conflict with the underlying preferences, or its operations introduce new assumptions that implicitly change the underlying preference order.

Theorem 19 (Combination) *If c, l, and d have disjoint support, $c \trianglerighteq d$, $l \trianglerighteq d$, and desire(c), then $c \wedge l \trianglerighteq d$.*

PROOF: By definition, $c \wedge l \trianglerighteq d$ holds iff $m[c \wedge l \wedge \neg d] \succsim m[\neg(c \wedge l) \wedge d]$ for all m. According to Lemma 8, the latter expression can be decomposed into the union of $m[\neg c \wedge l \wedge d]$, $m[c \wedge \neg l \wedge d]$, and $m[\neg c \wedge \neg l \wedge d]$. We establish the preference of $m[c \wedge l \wedge \neg d]$ over each of these. First,

$$m[c \wedge l \wedge \neg d] = m[l][c \wedge \neg d] \succsim m[l][\neg c \wedge d] = m[\neg c \wedge l \wedge d],$$

by $c \trianglerighteq d$, disjoint support, Lemma 7, and the uniqueness of $m[l]$. Similarly,

$$m[c \wedge l \wedge \neg d] = m[c][l \wedge \neg d] \succsim m[c][\neg l \wedge d] = m[c \wedge \neg l \wedge d],$$

by $l \trianglerighteq d$, disjoint support, Lemma 7, and the uniqueness of $m[c]$. Chaining this result with

$$m[c \wedge \neg l \wedge d] = m[\neg l \wedge d][c] \succsim m[\neg l \wedge d][\neg c] = m[\neg c \wedge \neg l \wedge d],$$

which holds by virtue of *desire*(c), disjoint support, Lemma 7, and the uniqueness of $m[\neg l \wedge d]$, yields $m[c \wedge l \wedge \neg d] \succsim m[\neg c \wedge \neg l \wedge d]$, the final condition we require. □

Theorem 20 (Conjunction of desires) *If c and c' have disjoint support, desire(c), and desire(c'), then desire$(c \wedge c')$.*

PROOF: We need to show that $m[c \wedge c']$ is preferred to any model in $m[c \wedge \neg c']$, $m[\neg c \wedge c']$, or $m[\neg c \wedge \neg c']$. First,

$$m[c \wedge c'] = m[c][c'] \succsim m[c][\neg c'] = m[c \wedge \neg c'],$$

by $desire(c')$, disjoint support, Lemma 7, and the uniqueness of $m[c]$. A symmetric argument with c and c' switched yields $m[c \wedge c'] \succsim m[\neg c \wedge c']$. Finally, let $m' \in m[c \wedge c']$ and $m'' \in m[\neg c \wedge \neg c']$. Note that $m \equiv m' \bmod c \wedge c'$, and thus by Lemma 5, $m[\neg c \wedge \neg c'] = m'[\neg c \wedge \neg c']$. Therefore, $m'' \in m'[\neg c \wedge \neg c']$, and thus by the first part of Theorem 17, $m' \succsim m''$. $\square$

Theorem 21 (Disjunction of desires) *If p and p' have disjoint support, $desire(p)$, and $desire(p')$, then $desire(p \vee p')$.*

PROOF: The disjunctive desire holds iff $m[p \vee p'] \succsim m[\neg p \wedge \neg p']$ holds for all m. Let $m' \in m[p \vee p']$ and $m'' \in m[\neg p \wedge \neg p']$. Note that $m \equiv m' \bmod p \vee p'$, and thus by Lemma 5 and disjoint support, $m[\neg p \wedge \neg p'] = m'[\neg p \wedge \neg p']$. Therefore, $m'' \in m'[\neg p \wedge \neg p']$, and so $m' \succsim m''$ by the first part of Theorem 17. $\square$

Theorems 20 and 21 provide conditions under which conjoining and disjoining desires produces more complex desires. However, their converses do not generally hold. Thus subgoaling on conjunctions and disjunctions in AND/OR search need not always produce *bona fide* desires. Viewed semantically, the subgoals may have undesirable properties ("side-effects") in addition to their relation to the compound desire. In general, preferences over complex sentences tell us little about preferences over their constituent parts.

Although one cannot usually justify common goal manipulations due to the many preference orders compatible with desires alone, these operations can sometimes be validated conditional on additional restrictions, such as assumptions of preferential independence given combinations of conditions.

5 Restricted relative desire

We have defined relative desire in terms of preferences over the set of all models $\mathcal{M}$, but in many cases the knowledge available to the planner or used in the planner's construction rules out the occurrence of some of these interpretations. That is, some logically possible models may be epistemically or constitutionally impossible. In such cases, we should not demand that the planner express preferences over the irrelevant logically possible models, but only require a preference order over the models that represent epistemically or constitutionally possible situations or outcomes.

To make the logic of relative desire more practical in this way, we restrict all the previous definitions to a set $M \subseteq \mathcal{M}$ representing the epistemically or constitutionally possible models. We define $m[p]^M$, the modifications of m by p restricted to M, by

$$m[p]^M \stackrel{\text{def}}{=} m[p] \cap M,$$

and define restricted relative desire among sentences as follows.

Definition 5 (Restricted relative desire) *We say that p is* desired over q when restricted to M, *written $p \trianglerighteq^M q$ and read briefly as "p over q in M," iff for each $m \in M$, $m' \in m[p \wedge \neg q]^M$, and $m'' \in m[\neg p \wedge q]^M$ we have $m' \succsim m''$. We say that p is* strictly desired over q when restricted to M, *written $p \triangleright^M q$, iff $p \trianglerighteq^M q$ but not $q \trianglerighteq^M p$.*

This definition characterizes restricted relative desire in exactly the same way as relative desire, except that one considers only models and modifications in M rather than $\mathcal{M}$. In other words, $p \trianglerighteq q$ just means $p \trianglerighteq^{\mathcal{M}} q$.

We further extend the preceding definition by defining relative desire restricted by sentences and theories. We write $p \trianglerighteq^r q$, read "p over q given r", to mean $p \trianglerighteq^{[\![r]\!]} q$, and $p \trianglerighteq^T q$ to mean $p \trianglerighteq^{[\![T]\!]} q$, where $[\![T]\!] = \bigcap_{r \in T} [\![r]\!]$. Thus $p \trianglerighteq q$ iff $p \trianglerighteq^{true} q$.

Theorem 22 (Strengthening) *If $\models r' \rightarrow r$, then $p \trianglerighteq^r q$ implies $p \trianglerighteq^{r'} q$.*

Proof: Suppose $\models r' \rightarrow r$ and $p \trianglerighteq^r q$. Let $m \in [\![r']\!]$, $m' \in m[p \wedge \neg q] \cap [\![r']\!]$, and $m'' \in m[\neg p \wedge q] \cap [\![r']\!]$. Since $\models r' \rightarrow r$, $[\![r']\!] \subseteq [\![r]\!]$, so $m \in [\![r]\!]$. Since $m[p \wedge \neg q] \succsim m[\neg p \wedge q]$ by hypothesis, we have $m' \succsim m''$. □

Corollary 23 (Arbitrary restrictions) *$p \trianglerighteq q$ iff $p \trianglerighteq^r q$ for every r.*

Theorem 24 (Incompatible restrictions) *If $\models p \rightarrow \neg r$ or $\models q \rightarrow \neg r$, then $p \trianglerighteq^r q$*

Proof: If $\models p \rightarrow \neg r$, then $m[p \wedge \neg q] \cap [\![r]\!] = \emptyset$, while if $\models q \rightarrow \neg r$, then $m[\neg p \wedge q] \cap [\![r]\!] = \emptyset$. In either case, the stated relation holds trivially. □

In particular, $p \trianglerighteq^{false} q$ for all p, q.

Lemma 25 (Restricted modifications) *If p and r have disjoint support, it follows that $m[p]^r \subseteq m[p \wedge r]$.*

PROOF: Assuming disjoint support, $support(p) \subseteq support(p \wedge r)$. If $m' \in m[p]^r$, then $m' \models p \wedge r$ and $m \equiv m' \bmod p \wedge r$, so $m' \in m[p \wedge r]$. □

Theorem 26 (Explicit restriction) *If $p \wedge r \trianglerighteq q \wedge r$ and r has support disjoint from p and from q, then $p \trianglerighteq^r q$.*

PROOF: Assume $p \wedge r \trianglerighteq q \wedge r$ and disjoint support. By simplifying the definition, we obtain $m[p \wedge \neg q \wedge r] \succsim m[\neg p \wedge q \wedge r]$ for each $m \in \mathcal{M}$. To show that $p \trianglerighteq^r q$, assume that $m \in [\![r]\!]$, $m' \in m[p \wedge \neg q]^r$, and $m'' \in m[\neg p \wedge q]^r$. By Lemma 25, $m[p \wedge \neg q]^r \subseteq m[p \wedge \neg q \wedge r]$ and $m[\neg p \wedge q]^r \subseteq m[\neg p \wedge q \wedge r]$, so $m' \succsim m''$, as desired. □

The converse does not hold. To see this, suppose that $p, q, r_1, r_2 \in \mathcal{A}$ and $r = r_1 \vee r_2$. The assumption that $p \trianglerighteq^r q$ amounts to requiring that $\{p, \neg q, r_1, r_2\} \succsim \{\neg p, q, r_1, r_2\}$, $\{p, \neg q, r_1, \neg r_2\} \succsim \{\neg p, q, r_1, \neg r_2\}$, and $\{p, \neg q, \neg r_1, \neg r_2\} \succsim \{\neg p, q, \neg r_1, \neg r_2\}$. If we have $\{\neg p, q, r_1, \neg r_2\} \succ \{p, \neg q, r_1, r_2\}$ in addition, then $p \wedge r \trianglerighteq q \wedge r$ does not hold.

Restricted relative desire captures part of the notion of *framing* defined in [7]. Rather than beginning by supposing a set of atomic sentences and their models, our treatment there follows decision theory by supposing a set Ω of possible *outcomes* and then representing these outcomes as vectors of attributes by means of a one-to-one *framing function* $\phi : \Omega \to \prod_{i=1}^n A_i$. The image $\phi(\Omega)$ of outcomes under the framing constitutes the set of possible attribute vectors. The logical formalization of relative desire presented here specializes the original semantics by representing outcomes by vectors of binary attributes rather than by vectors of arbitrary attributes, that is, choosing $\phi : \Omega \to \mathcal{M}$. Restricting relative desire to a set of models M then corresponds to choosing the framing so that $\phi(\Omega) = M$.

Framings and their corresponding restrictions on possible outcomes play important roles in reasoning because they can, in some cases, determine which sentences represent desires. In other words, the set of sentences representing desires can change as one changes knowledge or constitution to make different outcomes possible. For example, suppose that $\mathcal{A}$ consists of the two independent sentences: p, "I wear a raincoat", and q, "I stand in the rain." We assume 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. We then have the intuitive result that $\neg q$ represents a desire, but neither p nor $\neg p$ does. Yet p represents a desire if we assume q holds; that is, $p \trianglerighteq^q \neg p$. This relativity of desire occurs quite easily, as shown by the following result.

Theorem 27 (Relativity of desire) *If $p \trianglerighteq q$ does not hold, then $q \rhd^{r} p$ for some r.*

PROOF: Suppose that $p \trianglerighteq q$ does not hold. By Corollary 23, choose r so that $p \trianglerighteq^{r} q$ does not hold. We may then find $m \in [\![r]\!]$, $m' \in m[p \wedge \neg q]$, and $m'' \in m[\neg p \wedge q]$ so that $m'' \succ m'$. Let $r' = (\bigwedge m') \vee (\bigwedge m'')$, and verify that $q \rhd^{r'} p$, as desired. □

In particular, if p is not a desire, then $\neg p$ is sometimes a strict desire, and if neither p nor $\neg p$ are desires, then one can find different restrictions making each a strict desire.

6 Conclusion

We have presented a logic of relative desire in which desires have a nontrivial semantics in terms of preferences over models, and in which one may construct preferences corresponding to specific sets of desires. The definition of desires formalizes the intuition that goals are propositions that are preferred to their opposites, other things being equal. The logic extends our previous work [7] by considering relative desire between arbitrary conditions expressed in a propositional language, and we provided a collection of inference rules that support reasoning about relative desire. We showed that while this logic displays some intuitive properties, it also reveals that desirability sometimes depends on knowledge or the possible states of the world, and that some common and seemingly natural goal operations are not always valid. Designers of planning architectures wishing to justify the behavior of their systems must therefore either provide further constraints on the meaning of desires, find other means for expressing preference information, or justify unsound manipulations on heuristic grounds. This highlights the importance of developing more refined languages for specifying the objectives of planning agents.

Ordinary desires can depend on probability judgments as well as preferences. Unlike other approaches to reasoning about desirability (e.g., Jeffrey's [2]), our inference rules do not depend on probability distributions. We could strengthen these rules to take advantage of probabilistic information when available, but some situations call for an ability to reason about desires separate from beliefs.

Further work will focus on strengthening and expanding the set of inference rules. The numerous restrictions on the form and and interrelations among sentences limit the applicability of the initial inference rules presented here. We expect that some of these restrictions may be alleviated by applying less ambiguous model modification rules, perhaps based on minimality criteria from the theory of conditionals and belief revision.

The logic presented here constitutes part of a comprehensive decision-theoretic account of planning (see also [6]), and a more thorough treatment of the issue of goals and utilities is in preparation [8]. The formal treatment of desire also plays an important role in the framework of agent-oriented programming (AOP), proposed in [5]. AOP—a specialization of object-oriented programming—views agents as modifying their mental states as a result of informing one another, requesting information, and performing other kinds of communicative acts. In current AOP setups, the mental state of an agent consists of "motivation-free" components: beliefs, commitments, choices, and capabilities. The logic of relative desire positions us to enrich the mental state of agents, thus expanding the AOP framework. In general, we expect that much might be learned by developing planning architectures which combine desires and goals with other preferences in a manner faithful to the logic of relative desire.

Acknowledgement: Jon Doyle is supported by the USAF Rome Laboratory and DARPA under contract F30602-91-C-0018, and by National Institutes of Health Grant No. R01 LM04493 from the National Library of Medicine.

References

[1] J. Doyle. A model for deliberation, action, and introspection. AI-TR 581, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, Cambridge, MA, 1980.

[2] R. C. Jeffrey. *The Logic of Decision.* University of Chicago Press, Chicago, second edition, 1983.

[3] R. L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* John Wiley and Sons, New York, 1976.

[4] J. Marks. Introduction: On the need for theory of desire. In J. Marks, editor, *The Ways of Desire.* Precedent Publishing, Chicago, 1986.

[5] Y. Shoham. Agent-oriented programming. Technical Report STAN-CS-90-1335, Stanford University Computer Science Department, 1990.

[6] M. P. Wellman. *Formulation of Tradeoffs in Planning Under Uncertainty.* Pitman and Morgan Kaufmann, 1990.

[7] M. P. Wellman and J. Doyle. Preferential semantics for goals. In *Proceedings of the National Conference on Artificial Intelligence*, 1991.

[8] M. P. Wellman, J. Doyle, and T. Dean. Goals, preferences, and utilities: A reconciliation. in preparation, 1991.