On Decision-Theoretic Foundations for Defaults *

Ronen I. Brafman and **Nir Friedman**
Stanford University
Dept. of Computer Science
Stanford, CA 94305-2140
{brafman,nir}@cs.stanford.edu
http://robotics.stanford.edu/people/{brafman,nir}

Abstract

In recent years, considerable effort has gone into understanding default reasoning. Most of this effort concentrated on the question of entailment, i.e., what conclusions are warranted by a knowledge-base of defaults. Surprisingly, few works formally examine the general role of defaults. We argue that an examination of this role is necessary in order to understand defaults, and suggest a concrete role for defaults: Defaults simplify our decision-making process, allowing us to make fast, approximately optimal decisions by ignoring certain possible states. In order to formalize this approach, we examine decision making in the framework of *decision theory*. We use probability and utility to measure the impact of possible states on the decision-making process. We accept a default if it ignores states with small impact according to our measure. We motivate our choice of measures and show that the resulting formalization of defaults satisfies desired properties of defaults, namely *cumulative* reasoning. Finally, we compare our approach with Poole's decision-theoretic defaults, and show how both can be combined to form an attractive framework for reasoning about decisions.

We make numerous assumptions each day: the car will start, the road will not be blocked, there will be heavy traffic at 5pm, etc. Many of these assumptions are defeasible; we are willing to retract them given sufficient evidence. Humans naturally state defaults and draw conclusions from default information. Hence, defaults seem to play an important part in common-sense reasoning. To use such statements, however, we need a formal understanding of what defaults represent and what conclusions they admit.

The problem of default entailment—roughly, what conclusions we should draw from a knowledge-base of defaults—has attracted a great deal of attention. Many researchers attempt to find "context-free" patterns of default reasoning (e.g., [Kraus *et al.*, 1990]). As this research shows, much can be done in this approach. We claim, however, that the utility of this approach is limited; to gain a better understanding of defaults, we need to understand in what situations we should be willing to state a default.

Our main thesis is that an investigation of defaults should elaborate their *role* in the behavior of the reasoning agent. This role should allow us to examine when a default is appropriate in terms of its implications on the agent's overall performance. In this paper, we suggest a particular role for defaults and show how this role allows us to provide a semantics for defaults. Of course, we do not claim that this is the only role defaults can play.

In many applications, the end result of reasoning is a choice of actions. Usually, this choice is not optimal; there is too much uncertainty about the state of the world and the effects of actions to allow for an examination of all possibilities. We suggest that one role of defaults lies in simplifying our decision-making process by stating assumptions that reduce the space of examined possibilities. More precisely, we suggest that a default $\varphi \rightarrow \psi$ is a license to ignore $\neg\psi$ situations when our knowledge amounts to φ.

One particular suggestion that can be understood in this light is ϵ-semantics [Pearl, 1989]. In ϵ-semantics, we accept a default $\varphi \rightarrow \psi$ if given the knowledge φ, the *probability* of $\neg\psi$ is very small. This small probability of the $\neg\varphi$ states gives us a license to ignore them. Although probability plays an important part in our decisions, we claim that we should also examine the *utility* of our actions. For example, while most people think that it is highly unlikely that they will die next year, they also believe that they should not accept this as a default assumption in the context of a decision as to whether or not to buy life insurance. In this context, the stakes are too high to ignore this outcome, even though it is unlikely. We suggest that the license to ignore a set should be given based on its impact on our decision. To paraphrase this view, we should accept *Bird* $\rightarrow$ *Fly* if assuming that the bird flies cannot get us into too much trouble.

To formalize our intuitions we examine decision-making in the framework of *decision theory* [Luce and Raiffa, 1957]. Decision theory represents a decision problem using several components: a set of *possible states*, a *probability measure* over these sets, and a *utility* function that assigns to each action and state a numerical value.

*To appear in IJCAI'95.

Classical decision theory then uses the *expected utility* of an action as a measure of its "goodness".

In order to define defaults we need to understand when can we "safely ignore" a set of situations. When we ignore a set of situations consistent with our knowledge φ, our expected utility calculations will only approximate the expected utility of actions given φ. Such an approximation can lead to erroneous perception of the quality of actions, and consequently, to bad decisions. We suggest that a set of states can be safely ignored if a reasonably good action is chosen even when these states are ignored. Consequently, we consider a default $\varphi \rightarrow \psi$ to be "safe" if the action we choose when we consider only $(\varphi \wedge \psi)$-states is a good approximation (in terms of expected utility) of the action we would choose had we considered all φ-states. To implement this idea we propose a measure on sets of states that captures their impact on the outcome of the decision-making process. We accept the default $\varphi \rightarrow \psi$ when the measure of $\varphi \wedge \neg\psi$ is very small relative to that of φ. We will show that the proposed measure satisfies our stated desideratum.

Our measure takes into account two factors: the probability of the set and the utilities of actions on this set. If the probability of a set is small, then it seems that we can ignore it. However, if the utilities of actions on this set are extreme, as in the insurance example above, then we might not want to ignore it. On the other hand, if the utilities of all actions on the set are very close, then all actions look similar on this set, so we should focus on the differences among actions elsewhere.

The contribution of this paper is twofold: First, it advocates a more concrete approach to the study of defaults in which a specific role for defaults is required; with such a role we can gain a better understanding of the semantics, formal properties, and applications of defaults. Second, it proposes a particular role for defaults in our decision-making process and examines suitable formal semantics that fulfill this role. Thus, we can understand the implication of various properties of defaults in a concrete setting: we can examine how such properties affect the agent's decision-making process. Moreover, our semantics *grounds* defaults in a well-established theory—decision theory. Thus, we can use the tools provided by this theory when formalizing our intuitions about decision making. It also provides common ground with other work that shares these tools. In particular, we examine the relation between our defaults and statements such "if φ, than a is an optimal action" that have been studied by Poole [Poole, 1992]. We combine the two types of statements in one framework, leading to a rich knowledge representation language. Because Poole's work shares the fundamental notions of decision theory, we can integrate his approach into our framework in a semantically clean way. Finally, decision-theoretic defaults supply us with a method for compiling decision-theoretic information into a compact form. This compact form may allow for faster, albeit approximate, on-line decision making.

We are certainly not the first to note the importance of utility considerations in default reasoning. Similar intuitions were mentioned in many of the early works on default and defeasible reasoning (e.g., [McCarthy, 1980]). In particular, several works use expected utility consideration in evaluation of heuristic rules (e.g., [Langlotz *et al.*, 1986]). More recently, decision-theoretic foundations for defaults were advocated by Shoham (1987) and Doyle (1989). Doyle provides a formal analysis of "Pascal's wager" and shows how an assumption (the existence of God) can be justified in terms of utility. Finally, Poole (1992) examined a concrete notion of defaults that are grounded in terms of decision theory. Unlike previous works (with the exception of [Poole, 1992], see Section 3) we make decision theory the basis of a formal definition of defaults.

This paper is organized as follows. In Section 1, we review the basic framework of decision theory and relevant results in default reasoning. In Section 2, we formalize our notion of defaults. We start with a simple definition and show that while it captures our intuitions to some extent, it has some deficiencies. In particular, it fails to satisfy several basic desired properties of default reasoning. We then develop a stronger notion of defaults that does satisfy these desirable properties. In Section 3, we relate our suggestion to Poole's decision-theoretic defaults [Poole, 1992]. We show that while these two notions are quite different, they can be combined to create a framework for reasoning about decisions. We conclude with a discussion in Section 4.

1 Preliminaries

1.1 Decision Theory

We start by reviewing the basic setting of decision theory. (For more details, see [Luce and Raiffa, 1957].) Decision theory deals with decisions in the face of uncertainty. A *decision-theoretic context* is a tuple $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \Pr, U)$, where $\mathcal{S}$ is a set of possible *states* of the world before the decision is made; $\mathcal{O}$ is a set of possible *outcomes* of actions, i.e., states of the world after the decision is made and carried out; $\mathcal{A}$ is a set of possible *actions*, each one is a function from $\mathcal{S}$ to $\mathcal{O}$; Pr is a *probability measure* over $\mathcal{S}$ that captures (subjective) likelihood of each state; and U is a *utility function* that maps outcomes in $\mathcal{O}$ to real numbers, that quantify the desirability of outcomes. In the following discussion, we usually assume that $\mathcal{S}, \mathcal{O}$ and $\mathcal{A}$ are fixed, and do not mention them explicitly.

In a fixed decision-theoretic context the *expected utility* of an action a given evidence $E \subseteq \mathcal{S}$ is defined as

$$EU_{(\Pr,U)}(a|E) = \sum_{s \in E} \Pr(s|E) \cdot U(a(s)),$$

where $\Pr(s|E)$ is the conditional probability of s given the evidence E. Classical decision theory prescribes that given our assessment of a probability and utility measures and given evidence E, we should choose an action that maximizes expected utility, i.e., an action a such that $EU_{(\Pr,U)}(a|E) = \max_{a' \in \mathcal{A}} EU_{(\Pr,U)}(a'|E)$. We denote by $MEU_{(\Pr,U)}(E)$ the expected utility of the best action given E, i.e., $\max_{a' \in \mathcal{A}} EU_{(\Pr,U)}(a'|E)$, and by $mEU_{(\Pr,U)}(E)$ the expected utility of the worst action, i.e., $\min_{a' \in \mathcal{A}} EU_{(\Pr,U)}(a'|E)$. (From here on we omit the subscript $(\Pr, U)$ whenever it is clear from the context.)

We note that decision theory is only interested in the relative ordering of actions, given E, i.e., the relations between $EU(a|E)$ and $EU(a'|E)$. Since using a utility measure $U'(\cdot) = c_1 \cdot U(\cdot) + c_2$, for some $c_1 > 0$ and c_2, leads to the exact same conclusions, decision theory treats U and U' as equivalent.

Decision theory usually does not deal explicitly with how we describe events or actions. However, in our discussion of defaults we describe events using a logical language. We assume that there is a language $\mathcal{L}$ that is closed under the usual propositional connectives and a *truth-assignment* π that assigns to each state $s \in \mathcal{S}$ a subset of $\mathcal{L}$. Intuitively, $\pi(s)$ is the set of sentences that are true at s. We require that the following conditions hold:

- $\varphi \in \pi(s)$ if and only if $\neg\varphi \notin \pi(s)$.
- $\varphi \wedge \psi \in \pi(s)$ if and only if $\varphi \in \pi(s)$ and $\psi \in \pi(s)$.

From now on we will use $\Pr(\varphi)$ as an abbreviation of $\Pr(\{s|\varphi \in \pi(s)\})$.

1.2 Defaults

The study of default statements has a long tradition in artificial intelligence (see [Ginsberg, 1987; Gabbay *et al.*, 1993] for overviews). We denote by $\varphi \to \psi$ the statement "if φ then by default ψ". A typical example is the following statements: $Bird \to Flies$ and $Bird \wedge Penguin \to \neg Flies$. These two defaults state that birds typically fly, but penguins are exceptional and typically do not fly: default statements differ from material implication in that they allow for exceptions. Defaults are intuitively appealing and seem to provide a natural language for specifying common-sense knowledge. Formal understanding of defaults turns out to be quite elusive; there has been a great deal of discussion in the literature as to what the appropriate semantics of defaults should be. While there is little consensus on the semantics of defaults, there has been some consensus on reasonable "core" properties defaults. This core was suggested by Kraus, Lehmann and Magidor (1990) and consists of the following properties:

REF. $\varphi \to \varphi$ (Reflexivity).

LLE. If $\varphi \equiv \varphi'$, then from $\varphi \to \psi$ infer $\varphi' \to \psi$ (Left Logical Equivalence).

RW. If $\psi \Rightarrow \psi'$, then from $\varphi \to \psi$ infer $\varphi \to \psi'$ (Right Weakening).

CUT. From $\varphi \to \psi_1$ and $\varphi \wedge \psi_1 \to \psi_2$ infer $\varphi \to \psi_2$.

CM. From $\varphi \to \psi_1$ and $\varphi \to \psi_2$ infer $\varphi \wedge \psi_1 \to \psi_2$ (Cautious Monotonicity).

OR. From $\varphi_1 \to \psi$ and $\varphi_2 \to \psi$ infer $\varphi_1 \vee \varphi_2 \to \psi$.

REF states that φ is always a default conclusion of φ. LLE states that the syntactic form of the antecedent is irrelevant: logically equivalent antecedents have the same consequences. RW describes a similar property of the consequent: If ψ (logically) entails ψ', then we can deduce $\varphi \to \psi'$ from $\varphi \to \psi$. This allows us to combine default and logical reasoning. CM and CUT state that if ψ_1 is a default conclusion of φ, then ψ_2 is a default conclusion of φ if and only if it is a default conclusion of $\varphi \wedge \psi_1$: Discovering that ψ_1 holds (as would be expected, given the default) should not cause us to retract or add other default conclusions. OR states that we are allowed to reason by cases: If the same default conclusion follows from each of two antecedents, then it also follows from their disjunction.

Kraus, Lehmann and Magidor focus on *consequence relations*. A consequence relation captures a particular way we make assumptions. Given a pair of formulas φ and ψ, this relation determines whether we are willing to assume ψ given the knowledge φ. Formally, they define a consequence relation Cn to be the set of defaults, such that $\varphi \to \psi \in Cn$ if ψ is among the consequences of φ. Kraus, Lehmann and Magidor characterize *cumulative* reasoning by system **C**, composed of REF, LLE, RW, CM, and CUT, and *preferential* reasoning by system **P** that contains system **C** and OR. A consequence relation is *cumulative* (resp. *preferential*) if it satisfies system **C** (resp. system **P**), i.e., the set of defaults is closed under applications of these rules. They suggest that a "reasonable" consequence relation should be preferential. Furthermore they provide representation theorems for cumulative and preferential consequence relations using order relations over worlds. While we do not go into the motivation for these rules, they are accepted as reasonable "core" properties that nonmonotonic reasoning should satisfy.

Surprisingly, Pearl (1989) describes a probabilistic notion of defaults, ϵ-semantics, that leads to preferential consequence relations. Intuitively, ϵ-semantics accepts a default $\varphi \to \psi$ if $\Pr(\neg\psi|\varphi)$ is very small. Formally, to model "very small", ϵ-semantics examines behavior in the limit. A *parameterized probability distribution*[1] PPD is a family $= \{\Pr_n \mid n > 0\}$. Given a PPD, PP, the induced consequence relation is

$$Cn_\epsilon(PP) = \{\varphi \to \psi \mid \lim_{n\to\infty} \Pr_n(\neg\psi|\varphi) = 0\}.^2$$

Then it can be shown that:

Lemma 1.1: *[Goldszmidt* et al.*, 1993] Cn is a preferential consequence relation if and only if there is a PPD PP such that $Cn = Cn_\epsilon(PP)$.*

2 Decision-Theoretic Defaults

Our approach is based on the following idea: Given an appropriate measure of a set of states importance in the decision-making process, we can ignore those states of negligible importance. Thus, we will accept the default $\varphi \to \psi$ if the "importance" of $\varphi \wedge \neg\psi$ is very small in comparison to the importance of φ. In what follows, we investigate two definitions that try to capture this idea.

2.1 Basic Definition

One natural candidate is the maximal expected utility of a set. Suppose we know that we are in the set φ. Then we can write

$$EU(a|\varphi) = \Pr(\psi|\varphi)\cdot EU(a|\varphi\wedge\psi) + \Pr(\neg\psi|\varphi)\cdot EU(a|\varphi\wedge\neg\psi).$$

[1] Our presentation follows the formulation of [Goldszmidt *et al.*, 1993].

[2] To handle cases where $\Pr_n(\varphi) = 0$, we define $\Pr_n(\neg\psi|\varphi)$ to be 0 when $\Pr_n(\varphi) = 0$.

Thus, $\Pr(\neg\psi|\varphi) MEU(\varphi \wedge \neg\psi)$ is an upper bound on the contribution of $\neg\psi$ to the value of actions in φ. However, this may be misleading. For example, the expected utility of all actions on $\varphi \wedge \neg\psi$ might be the same high value. Intuitively, in this case $\neg\psi$ plays no role in determining what action is best on φ, yet $MEU(\varphi \wedge \neg\psi)$ is large. Moreover, as we noted above, any positive linear transformation of utilities (i.e., define $U'(\cdot) = c_1 \cdot U(\cdot) + c_2$ for some constants c_1 and c_2, s.t., $c_1 > 0$) should not change our conclusions. Yet, we can use such a transformation to blow up the maximum expected utility on sets. Therefore, instead of using MEU as the estimate, we use the following "normalized" measure:

$$\begin{aligned} G_{(\Pr,U)}(A) &=_{\text{def}} MEU(A) - mEU(A) \\ &= \max_{a,a' \in \mathcal{A}} (EU(a|A) - EU(a'|A)) \quad (1) \end{aligned}$$

(Again, we omit $(\Pr, U)$ when it is clear from the context.) We call $G(A)$ the *gain* of A, since it measures how much can be gained if we choose a good action instead of a bad one on A. It is easy to check that $\Pr(\neg\psi|\varphi) \cdot G(\varphi \wedge \neg\psi)$ bounds the potential loss incurred by ignoring $\neg\psi$ in the computation of expected utilities of actions:

$$|EU(a|\varphi) - \Pr(\psi|\varphi) \cdot EU(a|\varphi \wedge \psi)| \leq \Pr(\neg\psi|\varphi) \cdot G(\varphi \wedge \neg\psi).$$

However, we should remember that this error is relative to $G(\varphi)$, since we cannot do worse than $mEU(\varphi)$ nor better than $MEU(\varphi)$. This suggests that when we are willing to tolerate an error ratio of ϵ we can ignore $\neg\psi$ when we know φ if

$$\frac{\Pr(\neg\psi|\varphi) \cdot G(\varphi \wedge \neg\psi)}{G(\varphi)} \leq \epsilon$$

That is, $\Pr(\psi|\varphi) \cdot EU(a|\varphi \wedge \psi)$ is ϵ close to the actual expected utility on a when we know φ.

However, we usually do not want to fix an arbitrary ϵ. We overcome this problem by examining what happens in the limit when our threshold approaches 0. A *parameterized decision-theoretic context* (PDC) is a sequence $\{(\Pr_n, U_n) | n \geq 0\}$ of decision-theoretic contexts. Such a sequence describes our assessment of the decision problem when we successively lower the size of "ignorable" quantity. We define our first notion of defaults, which we will call "weak" for reasons that will become apparent below.

Definition 2.1 : A PDC $P = \{(\Pr_n, U_n) \mid n \geq 0\}$ (weakly) satisfies the default $\varphi \to \psi$, denoted $P \models_w \varphi \to \psi$, if

$$\lim_{n\to\infty} \frac{\Pr_n(\neg\psi|\varphi) \cdot G_n(\neg\psi \wedge \varphi)}{G_n(\varphi)} = 0, \qquad (2)$$

where G_n is an abbreviation for $G_{(\Pr_n, U_n)}$.[3] ∎

We define the consequence relation $Cn_w(P) = \{\varphi \to \psi \mid P \models_w \varphi \to \psi\}$.

Note that, according to this definition, we ignore $\neg\psi$ if the product of $\Pr(\neg\psi|\varphi)$ and the gain of $\varphi \wedge \neg\psi$ is small. In line with our intuitions, this definition weighs both the probability of the set, and the utility of actions on states in the set. It is also easy to see that this definition generalizes ϵ-semantics. If we choose utilities such that $G_n(A) = c$, for some constant c, for all non-empty sets A, then (2) becomes $\lim_{n\to\infty} \Pr_n(\neg\psi|\varphi) = 0$, which is equivalent to the definition of defaults in ϵ-semantics.[4] Thus, under certain choices of the utility function our definition becomes equivalent to ϵ-semantics.

Proposition 2.2: *There is a utility function U_ϵ, such that for each PPD $PP = \{\Pr_n \mid n \geq 0\}$, the PDC $P_{PP} = \{(\Pr_n, U_\epsilon) \mid \geq 0\}$ is such that $Cn_\epsilon(PP) = Cn_w(P_{PP})$.*

Above we stated the desideratum that defaults should not affect the quality of our decisions. Intuitively, an action is approximately optimal if, in the limit, its expected utility over E is almost as good as $MEU_n(E)$. Formally, we say that an action a is *approximately optimal* on a set E with respect to a PDC P, if

$$\lim_{n\to\infty} \frac{MEU_n(E) - EU_n(a|E)}{G_n(E)} = 0 \qquad (3)$$

Again, we must normalize by $G_n(E)$ to avoid being sensitive to positive linear transformations. We say that a default $\varphi \to \psi$ is *approximation safe* (with respect to P) if every approximately optimal action on $\varphi \wedge \psi$ is also approximately optimal on φ. This implies that choosing a good action on $\varphi \wedge \psi$ leads to a good action on φ.

Theorem 2.3: *If $P \models_w \varphi \to \psi$, then $\varphi \to \psi$ is approximation safe w.r.t. P.*

Definition 2.1 satisfies our stated criteria of approximation. However, the induced consequence relations, in general, are not cumulative. In particular, RW does not hold. Consider the following example, where we have two propositions p and q, and four equiprobable states. Utilities (for any n) of two actions a_1 and a_2 are defined according to this table:

	$p \wedge q$	$p \wedge \neg q$	$\neg p \wedge q$	$\neg p \wedge \neg q$
a_1	10	0	5	5
a_2	0	10	10	10

It is easy to check that $true \to \neg p$ is satisfied according to Definition 2.1, simply because $G(p) = 0$. However, $true \to (\neg p \vee \neg q)$ is not satisfied since $G(p \wedge q) = 10$. RW is violated because the gain of a set might be small, while the gain of (some of) its subsets might be very high. This phenomena occurs because of "canceling out", i.e., actions that are good on one subset are bad on another, and vice versa. In our example, a_1 and a_2 "cancel out" on p. When we examine the whole set, this phenomenon is undetectable, since we only examine the expected utilities of actions. It is easy to construct similar counterexamples to CUT, CM and OR.

This example shows that Definition 2.1 is quite weak. Before we discuss this issue, we examine what properties

[3] In order to handle defaults in situations where $G_n(A) = 0$ in a reasonable manner, we use throughout the paper the following definition: $\frac{x}{y} = 0$ whenever $x = y = 0$, and $\frac{x}{y}$ is infinitely large, i.e., unbounded, when $x > y = 0$.

[4] In fact, it suffices to require that U_n is such that for all non-empty sets A, $0 < c \leq G_n(A) \leq d$ for some constants c, d. This implies that $P \models_w \varphi \to \psi$ if and only if $\{\Pr_n\} \models_\epsilon \varphi \to \psi$.

are satisfied by this definition. We define the following weak variant of RW:

RW$_w$. If $\psi_1 \Rightarrow \psi_2$, then from $\varphi \rightarrow \psi_1$ and $\varphi \rightarrow (\psi_2 \Rightarrow \psi_1)$ infer $\varphi \rightarrow \psi_2$.

To understand the nature of this rule, we need to examine properties of G. As we noted above, if $B \subset A$, then $G(A)$ does not necessarily provide an upper bound on $G(B)$. This is an artifact of "canceling out", i.e., a big difference in the expected utility of actions on B is canceled out by their expected utility on $G(A \setminus B)$. But this implies that if $G(B)$ is much bigger than $G(A)$, then $G(A \setminus B)$ must also be big. In fact we can show that if $G(A)$ is "small", then $G(B)$ and $G(A \setminus B)$ must be of the same magnitude. Using this insight we can understand RW$_w$: From $\varphi \rightarrow \psi_1$ we infer that $G(\varphi \wedge \neg\psi_1)$ is small. The formulae $\varphi \wedge \neg\psi_1 \wedge \neg\psi_2$ and $\varphi \wedge \neg\psi_1 \wedge \psi_2$ form a disjoint partition of $\varphi \wedge \neg\psi_1$; if one of them is small, then the other is also. From $\varphi \rightarrow (\psi_2 \Rightarrow \psi_1)$ we conclude that $G(\varphi \wedge \neg\psi_1 \wedge \psi_2)$ is small, hence $G(\varphi \wedge \neg\psi_1 \wedge \neg\psi_2)$ is small. But, this is exactly the desired conclusion.

Similar reasoning leads to the following weak versions of CUT and CM:

CUT$_w$. From $\varphi \rightarrow \psi_1$, $\varphi \rightarrow \psi_1 \vee \psi_2$ and $\varphi \wedge \psi_1 \rightarrow \psi_2$, infer $\varphi \rightarrow \psi_2$.

CM$_w$. From $\varphi \rightarrow \psi_1$, $\varphi \rightarrow \psi_1 \vee \psi_2$ and $\varphi \rightarrow \psi_2$, infer $\varphi \wedge \psi_1 \rightarrow \psi_2$.

Let system $\mathbf{C}_w$ be the system containing REF, LLE, RW$_w$, CUT$_w$ and CM$_w$. We can then show:

Theorem 2.4: *If P is a PDC, then $Cn_w(P)$ satisfies system $\mathbf{C}_w$.*

It is unclear to us at this stage whether system $\mathbf{C}_w$ is complete, or whether there are other rules that hold for this definition. Notice that From CUT$_w$ and RW we can derive CUT, and from CM$_w$ and RW we can derive CM. Thus, the main difference between system $\mathbf{C}_w$ and system $\mathbf{C}$ is the weaker version of right weakening.

These results show that the most natural definition of defaults that satisfies our decision-theoretic desiderata (i.e., being approximation safe) has very weak properties. We consider the failure to satisfy properties of cumulative reasoning to be a serious one. Two properties of cumulative reasoning are especially important. The first is the AND property:

AND. From $\varphi \rightarrow \psi_1$ and $\varphi \rightarrow \psi_2$ infer $\varphi \rightarrow \psi_1 \wedge \psi_2$.

This property is derived from system $\mathbf{C}$ (see [Kraus *et al.*, 1990]). This property deals with *modularity* of assumptions. It states that if we can safely assume ψ_1 and also safely assume ψ_2, then we should be able to assume both. This property, however, is not guaranteed by Definition 2.1. The other property is CM. It states that if we happen to learn that some of our assumptions are true, we do not retract our previous assumptions. Suppose, for example, that I assume by default that my car will start, and that it is a sunny day. If, I then learn that the day is sunny, it seems intuitive that I should not need to retract my conclusions about the car. Again, this property is not guaranteed by Definition 2.1.

In general, we believe that properties of cumulative reasoning are indeed basic properties of any notion of defaults; if we do not satisfy cumulative reasoning, we must reexamine our assumptions whenever we have additional information, even if this information is consistent with our previous default conclusions. Such behavior seems undesirable. Thus, we would like to add the additional desideratum that accepted defaults are cumulative. This leads us to ask: is there a natural definition of decision-theoretic defaults that satisfy both desiderata?

2.2 Strong Definition

We have seen that our definition of defaults is "almost" cumulative, except that it does not satisfy RW. The problem was that even if $G(A)$ is small, it might be that $G(B)$, for some subset B of A, is very large. In other words, the measure $G(A)$ is not informative about the behavior of actions on subsets of A. To overcome this problem we introduce a more cautious measure of sets. We define

$$\Delta_{(\Pr,U)}(A) =_{\text{def}} \max_{B \subseteq A} \Pr(B|A) \cdot G_{(\Pr,U)}(B) \qquad (4)$$

for non-empty A, and define $\Delta_{(\Pr,U)}(\emptyset) = 0$. (Again, we omit the subscript when it is clear from the context.) It is easy to check that if $B \subseteq A$, then $\Pr(B|A) \cdot \Delta(B) \leq \Delta(A)$. Thus, $\Delta(A)$ is more informative about the behavior of subsets of A than $G(A)$. In particular, if for some ϵ,

$$\frac{\Pr(\neg\psi|\varphi) \cdot \Delta(\neg\psi \wedge \varphi)}{G(\varphi)} \leq \epsilon,$$

then we can conclude that

$$\frac{\Pr(\neg\psi'|\varphi) \cdot G(\neg\psi' \wedge \varphi)}{G(\varphi)} \leq \epsilon$$

for all ψ' such that $\psi \Rightarrow \psi'$. This suggests that the following definition satisfies our desiderata.

Definition 2.5: A PDC $P = \{(\Pr_n, U_n) \mid n \geq 0\}$ (strongly) satisfies the default $\varphi \rightarrow \psi$, denoted $P \models_s \varphi \rightarrow \psi$, if

$$\lim_{n \rightarrow \infty} \frac{\Pr_n(\neg\psi|\varphi) \cdot \Delta_n(\neg\psi \wedge \varphi)}{G_n(\varphi)} = 0, \qquad (5)$$

where Δ_n is an abbreviation for $\Delta_{(\Pr_n,U_n)}$. ∎

We define the (strong) consequence relation of P as $Cn_s(P) =_{\text{def}} \{\varphi \rightarrow \psi | P \models_s \varphi \rightarrow \psi\}$. It is easy to verify that this definition of defaults is indeed more restrictive than Definition 2.1.

Proposition 2.6: *Let P be a PDC. If $P \models_s \varphi \rightarrow \psi$, then $P \models_w \varphi \rightarrow \psi$.*

An immediate corollary is that if $\varphi \rightarrow \psi$ is strongly satisfied by P then it is approximation safe with respect to P. Moreover, we can show that this notion of defaults satisfies cumulative reasoning.

Theorem 2.7: *If P is a PDC, then $Cn_s(P)$ satisfies system $\mathbf{C}$.*

We conjecture that system **C** is complete with respect to the class of all PDCs, i.e., a consequence relation Cn is cumulative if and only if there exists a PDC P, such that $Cn = Cn_s(P)$.

This result shows that Definition 2.5 satisfies our desiderata using a natural decision-theoretic construction. It might seem that our definition of Δ is somewhat arbitrary. Indeed, as we show in the full version of this paper, similar properties are satisfied by other measures as well. Roughly, we show that Δ' satisfies Proposition 2.6 if and only if $\Delta'_n(A)$ provides an upper bound, in a certain precise sense, on $G_n(A)$, and Δ' satisfies Theorem 2.7 if and only if $\Delta(A)$ provides an upper-bound on $\Pr(B|A) \cdot \Delta(B)$ for $B \subseteq A$. We claim that Δ is the most natural member of this family.

2.3 The OR Rule

The last section showed how to obtain cumulative reasoning in our framework. Recall that preferential reasoning is defined to be cumulative reasoning combined with the OR rule. Most accepted semantics of defaults, in particular preferential structures and ϵ-semantics, lead to preferential consequence relations. Is OR satisfied in the two approaches we described? As the following example shows, this is not necessarily the case.

Example 2.8: Consider the following scenario. The agent is contemplating two possible investments. He can either buy the stocks of company A, an oil producer, or those of company B, a plastic manufacturer. The success of either investment is greatly dependent on changes in the price of oil. If oil prices rise, company A's profits will increase. However, since plastic is an oil by-product, the cost of raw material for company B will rise, and its profits will decline. On the other hand, if oil prices decline, company A's profits will decline, and company B's profits will increase. This situation is complicated by news of a technological break-through in oil refinement. This technology is expected to decrease the cost of oil refinement, reducing the costs for both companies. But it will have a more dramatic effect on company B by improving the quality of its raw material. However, this technology is still in early stages of development, and is not likely to have any effect on the market in the next few years.

These considerations are captured by the following (parameterized) decision theoretic setting.

	$O^+ \wedge \neg T$	$O^+ \wedge T$	$O^- \wedge T$	$O^- \wedge \neg T$
$\Pr_n$	$1/2 - 1/n$	$1/n$	$1/n$	$1/2 - 1/n$
A	1	6	4	-1
B	-1	9	11	1

Suppose the agent knows that oil prices will rise. Then, he can ignore the possible emergence of new technologies. To see this, $G(O^+) = 2 + 1/n$, and $G(O^+ \wedge T) = 3/n$, thus we would accept $O^+ \rightarrow \neg T$. Similar consideration show that if the agent knows that oil prices will fall, he can also ignore the new technology, i.e., $O^- \rightarrow \neg T$. What happens when the agent does not know whether oil price will rise or fall? In that case, he cannot ignore the possibility of new technology. Without knowledge about the direction of oil prices, investing in A or in B is more or less the same, except when the new technology arrives. In that situation the plastic industry is clearly a better choice. This type of consideration, which is secondary when the agent has more knowledge, becomes a major consideration in the decision *without* that knowledge. Technically, we have that $G(O^+ \vee O^-) = G(T) = 10/n$, thus we cannot accept the default $O^+ \vee O^- \rightarrow \neg T$. And if the agent accepts this default, he is likely to make the wrong choice, i.e., buy into the oil company. ∎

In retrospect, it is not too surprising that OR is not satisfied in our system. The essence of OR is reasoning by cases: If when φ_1 holds we can assume ψ, and when φ_2 holds we can assume ψ, then we should also assume ψ when we know that one of these cases is true. However, as noticed by Kraus, Lehmann and Magidor, this rule might be inappropriate when we read the antecedent of the default as "I *only* know φ" (which is basically how we interpret this default): "I *only* know φ" is not equivalent to "I *only* know $\varphi \wedge \psi$" or "I *only* know $\varphi \wedge \neg\psi$."

3 Poole's decision-theoretic defaults

Poole (1992) introduces a semantics for defaults that is also based on decision theory. His motivation is similar to our own, yet his proposal is very different. We now briefly review his semantics. A default in Poole's system has the form $\varphi \leadsto a$, and reads "Given evidence φ, do action a". This default caches information about the best action to perform when we get evidence φ. Such a default is accepted in a decision-theoretic context $(\Pr, U)$ if a maximizes the expected utility over φ, that is

$$EU(a|\varphi) = MEU(\varphi). \tag{6}$$

Poole argues that this definition naturally captures many real-life defaults. He gives examples of default statements that conclude what action to perform, such as "if you are in Vancouver in November, carry an umbrella". He shows that his semantics satisfies several desirable criteria, such as non-monotonicity, specificity, and ignoring irrelevant information.

Poole would ultimately want his semantics to capture regular defaults, such as "birds typically fly." However, these defaults have formulas as their conclusions, not actions. Poole attempts to overcome this problem using the following idea: With each proposition p, he associates three actions: p^t, p^f, p^u. These actions stand for: assume that p is true, assume that p is false, and do not make assumptions on p, respectively. Poole then represents defaults such as "birds typically fly" as $Bird \leadsto Fly^t$. He shows that under certain (rather strong) assumptions on the utility of these actions, he can give accepting conditions for defaults in terms of Pr and U. Poole's solution forces us to examine utilities of actions of a specific form – making assumptions. It seems to us that unless we have a good model of how making assumptions affects the choice of "real" actions (i.e., actions in the world), it is quite difficult to assess their utility. Moreover, it is unclear whether such a model will satisfy requirements of Poole's analysis. Our approach to defaults circumvents these problems by examining the utility of the actual actions available to us when we face the decision. We believe this approach is more natural. In any particular context we are facing a choice between several concrete decisions. The context describes the possible outcomes these decisions can lead to and their resulting utilities.

In spite of this criticism, we believe that defaults of the form $\varphi \leadsto a$ are useful, and suggest that Poole's defaults can be combined with our notion of defaults. This leads to a system where we can state defaults about actions to perform, as well as assumptions that can be made. We now outline the synthesis of Poole's defaults and our system into a joint framework.

Recall that Poole's original semantics for $\varphi \leadsto a$ is that $EU(a|\varphi) = MEU(\varphi)$. We want to define an similar definition in terms of PDCs. Instead of stating that a is the best action, given φ, we state that a is a safe approximation. Formally, a PDC P satisfies $\varphi \leadsto a$ if (3) holds. This definition is very much in the spirit of Poole's original one, and we can show that the same properties are satisfied. Moreover, there are interesting interactions between Poole's defaults and ours.

Theorem 3.1: *The following rules of inference are valid for $\rightarrow$ and $\leadsto$:*

- *If $\varphi \equiv \varphi'$, then from $\varphi \leadsto a$ infer $\varphi' \leadsto a$.*
- *From $\varphi \wedge \psi \leadsto a$ and $\varphi \wedge \neg\psi \leadsto a$ infer $\varphi \leadsto a$.*
- *From $\varphi \rightarrow \psi$ and $\varphi \wedge \psi \leadsto a$ infer $\varphi \leadsto a$.*
- *From $\varphi \rightarrow \psi$ and $\varphi \leadsto a$ infer $\varphi \wedge \psi \leadsto a$.*
- *From $\varphi \rightarrow$ false infer $\varphi \leadsto a$.*

This theorem shows that our definition of Poole's default satisfies the *sure thing principle* [Savage, 1954]: if a is a good action when I know $\varphi \wedge \psi$, and a good action when I know $\varphi \wedge \neg\psi$, then it is a good action when I just know φ. Moreover, the third and fourth properties show the direct relation between our defaults and approximately good actions: if $\varphi \rightarrow \psi$, then a good action when I know φ is also good when I know $\varphi \wedge \psi$, and vice versa. We believe that the combination of both types of default is useful. In future work we intend to apply this logical framework in applications to reasoning about decisions and knowledge compilation.

4 Discussion

Our starting point was the thesis that knowledge representation and reasoning methodologies are better understood in terms of their role in determining the behavior of agents. Once we have establish this role, we can gain a better understanding of the methodology in question. We examined one particular role of default reasoning. Focusing on this role helped us to determine desiderata that defaults should satisfy and to derive decision-theoretic semantics for defaults that meet these desiderata. Given this role, we can motivate what conclusions are entailed from a knowledge-base of defaults. But more importantly, providing a role for defaults is the first step toward understanding what defaults the knowledge-base should contain in the first place. As our approach suggests, the content of the knowledge-base depends on the specific context of the agent: his beliefs (i.e., probability) his goals (i.e., utility), and the actions available to him.

Our semantics *grounds* defaults in decision-theoretic contexts. Intuitively, this is because we use well-understood notions (i.e., probability and utility) instead of abstract ones (e.g., preferential structures). This choice allows us to relate defaults to other forms of knowledge. In this paper we examined one candidate, Poole's default actions. We believe that knowledge is not, in general, homogeneous. It is composed of various types of statements, and clearly there are interactions between these statements. Grounding these different types of statements in a common basis allows us to understand these interactions. In our case, the interactions between Poole's defaults and ours described in Theorem 3.1 are not arbitrary; they are a consequence of the semantics of both defaults in terms of decision-theoretic contexts.

Our definitions rely on PDCs – sequences of decision-theoretic contexts. These structures, which may not appear intuitive at first sight, should be understood as a mathematical idealization. This idealization allows us to talk about very small quantities or very big quantities, and in particular, the quotient ϵ, without committing to a particular value. This point highlights an important problem in nonmonotonic reasoning as well as probabilistic reasoning: what is a an acceptable notion of approximation? It is clear that setting a fixed threshold value is a crude way of defining approximation. Similarly, the use of limits is also quite crude. For example, we do not examine the rate of convergence nor do we provide a methodology for obtaining these sequences.

Of course, in real applications, we can often set a threshold value, below which things are considered small enough to be ignored. Once we fix this threshold we accept a default when the expression in (2) (or (5)) is smaller than this threshold. This definition approximates the notions we examined here. In particular, it does not satisfy the inference rules we describe. However, we can reason using these inference rules and get conclusions that might violate the fixed error margin. This provides a way of getting "fast and dirty" conclusions. Such an approach has been applied in the in ϵ-semantics literature, and recent work [Darwiche and Goldszmidt, 1994] indicate that such approximations might be quite useful. A possible avenue of future research is to use this method in knowledge-compilation of decision-theoretic information [Henrion *et al.*, 1991]. Roughly, in this method, off-line computation will generate a set of defaults using some parameter ϵ. These defaults (and their logical consequences) will be used at run-time to ignore various possibilities, hence reducing the amount of time spent in evaluating possible actions. As with any type of approximation, there is a tradeoff between the quality of the inference made (decision in this case) and the time spent on making this inference. Decision-theoretic defaults can be viewed as summarizing the information encoded in the underlying decision-theoretic context and may allow for faster on-line computations.

Our analysis is based on *static* or "one-shot" decision theory. Recently, there has been much work on decision making in dynamic environments (e.g., Markov Decision Processes [Puterman, 1994]). The notion of expected utility in these models is somewhat more complicated.

However, similar considerations of probability and utility apply when attempt to ignore various possibilities, i.e., we would like to ignore a possibility if it has small impact on the quality of actions we later choose. We intend to examine notions similar to default assumptions in the framework of Markov Decision Processes and to use these result to provide fast and approximately optimal planning in this setting.

Finally, we note that the approach we examine in this paper is not the only one for justifying defaults. In particular, several recent works [Pearl, 1993; Boutilier, 1994] examine approaches to *qualitative decision theory*. Roughly, these are analogues to decision theory where defaults play the role of probabilities, and analogues of utility, and expected utility (i.e., a combination rule) are suggested. All of these approaches are *descriptive* in that they espouse a particular procedure for decision-making. We believe that it is important to understand the *normative* foundations of such qualitative decision theory. This involves finding reasonable "rationality postulates" (in the sense of Savage's (1954) normative foundation for decision theory) that characterize these qualitative decision procedures. Initial results in this spirit appear in [Brafman and Tennenholtz, 1994], although in a somewhat different context. Such results should help us understand the consequences of adopting a specific representation for decision making under uncertainty. This type of investigation, which we are currently undertaking, should elucidate the tradeoffs between qualitative representations and quantitative representations.

Acknowledgements

The authors are grateful to Craig Boutilier, Jon Doyle, Hector Geffner, Moises Goldszmidt, James Kittock, Daphne Koller, Yoav Shoham, the anonymous referees and particularly Joe Halpern, for comments on drafts of this paper and useful discussions relating to this work. The first author was supported in part by ARPA grant AF F 49620-94-1-0090 and NSF grant IRI-9220645. The second author was supported in part by the Air Force Office of Scientific Research (AFSC), under Contract F49620-91-C-0080.

References

[Adams, 1975] E. Adams. *The Logic of Conditionals.* D. Reidel, Dordrecht, 1975.

[Boutilier, 1994] C. Boutilier. Toward a logic for qualitative decision theory. In *Principles of Knowledge Representation and Reasoning: Proc. 4th International Conference (KR '94)*, pp. 75–86. 1994.

[Brafman and Tennenholtz, 1994] R. I. Brafman and M. Tennenholtz. Belief ascription. In *Principles of Knowledge Representation and Reasoning: Proc. 4th International Conference (KR '94)*, pp. 87–98. 1994.

[Darwiche and Goldszmidt, 1994] A. Darwiche and M. Goldszmidt. On the relation between kappa calculus and probabilistic reasoning. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI '94)*, pp. 145–153. 1994.

[Doyle, 1989] J. Doyle. Constructive belief and rational representation. *Computational Intelligence*, 5:1–11, 1989.

[Gabbay *et al.*, 1993] D. M. Gabbay, C. J. Hogger, and J. A. Robinson, editors. *Nonmonotonic Reasoning and Uncertain Reasoning*, volume 3 of *Handbook of Logic in Artificial Intelligence and Logic Programming.* Oxford University Press, Oxford, U.K., 1993.

[Ginsberg, 1987] M. L. Ginsberg, editor. *Readings in Nonmonotonic Reasoning.* Morgan Kaufmann, San Francisco, Calif., 1987.

[Goldszmidt *et al.*, 1993] M. Goldszmidt, P. Morris, and J. Pearl. A maximum entropy approach to nonmonotonic reasoning. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 15(3):220–232, 1993.

[Henrion *et al.*, 1991] M. Henrion, J. S. Breese, and E. J. Horvitz. Decision analysis and expert systems. *AI Magazine*, 12:64–91, 1991.

[Kraus *et al.*, 1990] S. Kraus, D. J. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[Langlotz *et al.*, 1986] C. P. Langlotz, E. H. Shortliffe, and L. M. Fagan. Using decision theory to justify heuristics. In *Proc. National Conference on Artificial Intelligence (AAAI '86)*, pp. 215–219. 1986.

[Luce and Raiffa, 1957] R. D. Luce and H. Raiffa. *Games and Decisions.* Wiley, New York, 1957.

[McCarthy, 1980] J. McCarthy. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980.

[Pearl, 1989] J. Pearl. Probabilistic semantics for nonmonotonic reasoning: A survey. In *Proc. 1st International Conference on Principles of Knowledge Representation and Reasoning (KR '89)*, pp. 505–516. 1989.

[Pearl, 1993] J. Pearl. From conditional oughts to qualitative decision theory. In *Proc. 9th Conference on Uncertainty in Artificial Intelligence (UAI '93)*, pp. 12–20. 1993.

[Poole, 1992] D. Poole. Decision theoretic defaults. In *Proc. 8th Annual Conference on Uncertainty Artificial Intelligence (UAI '92)*, pp. 190–197. 1992.

[Puterman, 1994] M. Puterman. *Markov Decision Processes.* Wiley, New York, 1994.

[Savage, 1954] L. J. Savage. *Foundations of Statistics.* John Wiley & Sons, New York, 1954.

[Shoham, 1987] Y. Shoham. Nonmonotonic logics: meaning and utility. In *Proc. 10th International Joint Conference on Artificial Intelligence (IJCAI '87)*, pp. 388–393. 1987.