Defeasible Modes of Inference: A Preferential Perspective

Katarina Britz and **Ivan Varzinczak**
Centre for Artificial Intelligence Research
CSIR Meraka Institute and UKZN, South Africa
{arina.britz,ivan.varzinczak}@meraka.org.za

Abstract

Historically, approaches to defeasible reasoning have been concerned mostly with one aspect of defeasibility, viz. that of arguments, in which the focus is on normality of the *premise*. In this paper we are interested in another aspect of defeasibility, namely that of defeasible *modes* of reasoning. We do this by adopting a preferential modal semantics that we defined in previous work and which allows us to refer to the relative normality of accessible worlds. This leads us to define preferential versions of the traditional notions of knowledge, beliefs, obligations and actions, to name a few, as studied in modal logics. The resulting preferential modal logics make it possible to capture, and reason with, aspects of defeasibility heretofore beyond the reach of modal formalisms.

Introduction and Motivation

Defeasible reasoning, as traditionally studied in the literature on non-monotonic reasoning, has focused mostly on one aspect of defeasibility, namely that of arguments. Such is the case, for instance, in the well-known KLM approach (Kraus, Lehmann, and Magidor 1990; Lehmann and Magidor 1992), in which (propositional) defeasible consequence relations $\mathrel{|\!\sim}$ are studied. In this setting, the meaning of a defeasible statement (or a 'conditional', as it is sometimes referred to) of the form $\alpha \mathrel{|\!\sim} \beta$ is that "all normal α-worlds are β-worlds", leaving it open for α-worlds that are, in a sense, exceptional not to satisfy β. With the theory that has been developed around this notion it becomes possible to cope with exceptionality when performing reasoning.

There are of course many other appealing and equally useful aspects of defeasibility besides that of arguments. These include notions such as typicality (Giordano et al. 2009; Booth, Meyer, and Varzinczak 2012), concerned with the most typical cases or situations (or even the most typical representatives of a class), and belief plausibility (Baltag and Smets 2008), which relates to the most plausible epistemic possibilities held by an agent, amongst others. It turns out that with KLM-style defeasible statements one cannot capture these aspects of defeasibility. This has to do partly with the syntactic restrictions imposed on $\mathrel{|\!\sim}$, namely no nesting of conditionals, but, more fundamentally, it relates to where and how the notion of normality is used in such statements. Indeed, in a KLM defeasible statement $\alpha \mathrel{|\!\sim} \beta$, the normality spotlight is somewhat put on α, as though normality was a property of the premise and not of the conclusion. Whether the β-worlds are normal or not plays no role in the reasoning that is carried out. Furthermore, normality is assumed to be a property of the premise as a whole, and not of its constituents. Technically this means one cannot refer directly to normality of a sentence in the scope of logical operators. This is also the case in recent extensions of the KLM approach to logics that are more expressive than the propositional one (Britz, Heidema, and Meyer 2008; Britz, Meyer, and Varzinczak 2011a; 2011b).

In this paper we are interested in aspects of defeasibility related to the aforementioned idea of beliefs that are expressed in terms of most plausible accessible worlds. We investigate a more general notion which we refer to as defeasible *modes* of inference. These amount to preferential versions of the traditional notions of knowledge, beliefs, obligations and actions, to name a few, as studied in modal logics. For instance, in an action context, one may want to state that the outcome of a given action a is usually α, i.e., in the most normal situations resulting from the action's execution, α holds. This is notably different from saying that in the most normal worlds, the result of performing the action a is *always* α, i.e., stating $\top \mathrel{|\!\sim} \Box_a \alpha$ in Britz et al.'s (2011a) modal extension to preferential reasoning.

To give a more concrete example, one thing is to say that in any normal situation, a head-on collision at high speed results in a situation where there are fatalities, whereas another one is to say that in any situation, a head-on collision at high speed results in a situation in which there normally are fatalities. Here we are interested in the formalization of the latter type of statement, where it becomes important to shift the notion of normality from the antecedent of an inference to the effect of an action, and, importantly, use it in the scope of other logical constructors.

The importance of defeasibility in specific modes of reasoning is also illustrated by the following example. Although one may envisage a situation where the velocity of a sub-atomic particle in a vacuum is greater than c (the speed of light in a vacuum), it is in a sense known that c is the highest possible speed. Moreover, one can derive factual consequences of this scientific theory that also will be 'known'. This venturous version of knowledge, which patently dif-

fers from belief, provides for a more fine grained notion of knowledge that may turn out to be wrong, i.e., that is defeasible, but that is not of the same nature as suppositions or beliefs. Our proposal is not aimed at challenging the position of knowledge as indefeasible, justified true belief (Gettier 1963; Lehrer and Paxson 1969), but rather provides an extension to epistemic modal logics to allow for reasoning with a modality that we shall, argueably for lack of a more suitable term, refer to as "defeasible knowledge".

Our third example concerns obligations and weaker versions thereof. There is a subtle difference between stating that, in any normal situation, one ought to tell the truth, and stating that, in any situation, it is one's normal duty to tell the truth. In the latter the normality of the current situation is immaterial, whereas in the former it determines the truth of the statement. Therefore, the shift in focus is again from normality of the present world, to relative normality of accessible worlds.

Scenarios such as the ones depicted above require an ability to talk about the normality of effects of an action, normality of knowledge or obligations, and so on. While existing modal treatments of preferential reasoning can express preferential semantics syntactically as modalities (Boutilier 1994; Giordano et al. 2005; Britz, Heidema, and Labuschagne 2009), they do not suffice to express defeasible modes of inference as described above.

In this paper we fill this gap by introducing (non-standard) modal operators allowing us to talk about relative normality in accessible worlds. With our defeasible versions of modalities, we can make statements of the form "α holds in all of the normal accessible worlds", thereby capturing defeasibility of what is 'expected' in target worlds. Such a notion of defeasibility in a modality meets a variety of applications in Artificial Intelligence, ranging from reasoning about actions to deontic and epistemic reasoning. For instance, we define a defeasible-action operator allowing us to make statements of the form $\mathrel{\sim\!\!\Box}_a\alpha$, which we read as "$\alpha$ is a normal necessary effect of a", and we define defeasible-knowledge operators with which one can state formulas such as $\mathrel{\sim\!\!\Box}_A\alpha$, read as "agent A knows that normally α".

These operators are defined within the context of a general preferential modal semantics (Britz, Meyer, and Varzinczak 2011a). The relative normality of a given world in a Kripke model is determined by a preference order on *states*, serving as place holders for pointed Kripke models. In contrast with the plausibility models of Baltag and Smets (2008), our order on states does not define an agent's knowledge or beliefs. Rather, it is part of the semantics of the background ontology described by the theory or knowledge base at hand. As such, it informs the meaning of defeasible actions, which can fail in their outcome, or defeasible knowledge, which may not hold in exceptional accessible worlds, in that it alters the classical semantics of these modalities. This allows for the definition of a family of modal logics in which defeasible modes of inference can be expressed, and which can be integrated with existing non-monotonic modal logics.

The remainder of the present paper is structured as follows: In the next section we set up the modal notation of the paper and we recall the preferential semantics for modal logics that we shall use throughout this paper. Following that we present a logic enriched with defeasible modalities allowing for the formalization of defeasible versions of e.g. knowledge and actions, which we illustrate with examples in the following section. After a discussion of, and comparison with related work, we conclude with directions for further investigations.

Preliminaries

We commence by providing the required background for the rest of this work. First we recapitulate basic notions from modal logics and set up the notation we shall use.

Modal Logic

We work in a set of *atomic propositions* $\mathcal{P}$, using the logical connectives $\wedge$ (conjunction), $\neg$ (negation), and a set of modal operators $\Box_i$, $1 \leq i \leq n$. (In later sections we shall adopt a richer language.) Here we suppose that the underlying multimodal logic is independently axiomatized (i.e., the logic is a fusion and there is no interaction between the modal operators (Kracht and Wolter 1991)). Propositions are denoted by $p, q, \ldots$, and formulas by $\alpha, \beta, \ldots$, constructed in the usual way according to the rule:

$$\alpha ::= p \mid \neg\alpha \mid \alpha \wedge \alpha \mid \Box_i\alpha$$

All the other truth functional connectives ($\vee$, $\rightarrow$, $\leftrightarrow$, $\ldots$) are defined in terms of $\neg$ and $\wedge$ in the usual way. Given $\Box_i$, $1 \leq i \leq n$, with $\Diamond_i$ we denote its *dual* modal operator, i.e., for any α, $\Diamond_i\alpha \equiv_{\text{def}} \neg\Box_i\neg\alpha$. We use $\top$ as an abbreviation for $p \vee \neg p$, and $\bot$ for $p \wedge \neg p$, for some $p \in \mathcal{P}$.

With $\mathcal{L}$ we denote the set of all formulas of the modal language. The semantics is the standard possible-worlds one:

Definition 1 (Kripke Model) *A Kripke model is a tuple $\mathscr{M} = \langle W, R, V\rangle$ where W is a set of possible worlds, $R = \langle R_1, \ldots, R_n\rangle$, where each $R_i \subseteq W \times W$ is an accessibility relation on W, $1 \leq i \leq n$, and $V : W \times \mathcal{P} \longrightarrow \{0, 1\}$ is a valuation function.*

Figure 1 depicts two examples of Kripke models for $\mathcal{P} = \{p, q\}$. (In our pictorial representations of Kripke models we shall use $p \in w$ to mean that $V(w, p) = 1$ and $p \notin w$ to mean $V(w, p) = 0$.)

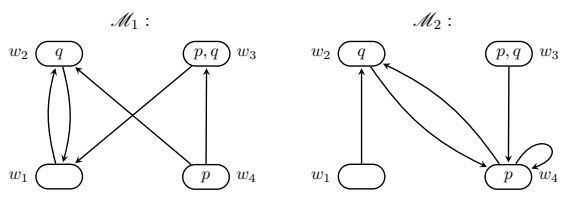

Figure 1: Examples of Kripke models.

Definition 2 (Satisfaction) *Let $\mathscr{M} = \langle W, R, V\rangle$, $w \in W$:*

- $\mathscr{M}, w \nVdash \bot$;
- $\mathscr{M}, w \Vdash p$ *if and only if* $V(w, p) = 1$;
- $\mathscr{M}, w \Vdash \neg\alpha$ *if and only if* $\mathscr{M}, w \nVdash \alpha$;

- $\mathscr{M}, w \Vdash \alpha \wedge \beta$ *if and only if* $\mathscr{M}, w \Vdash \alpha$ *and* $\mathscr{M}, w \Vdash \beta$*;*
- $\mathscr{M}, w \Vdash \Box_i \alpha$ *if and only if* $\mathscr{M}, w' \Vdash \alpha$ *for all* w' *such that* $(w, w') \in R_i$.

Given $\alpha \in \mathcal{L}$ and $\mathscr{M} = \langle W, R, V \rangle$, we say that $\mathscr{M}$ *satisfies* α if there is at least one $w \in W$ such that $\mathscr{M}, w \Vdash \alpha$. We say that $\mathscr{M}$ is a *model of* α (alias α is *true* in $\mathscr{M}$) if and only if $\mathscr{M}, w \Vdash \alpha$ for every $w \in W$. Given a class of models $\mathcal{M}$, we say that α is *valid* in $\mathcal{M}$ if every $\mathscr{M} \in \mathcal{M}$ is a model of α.

As we have just alluded to, among all possible models, one may want to choose some with specific properties to work with. This defines a *class of models*. A class of models $\mathcal{M}$ can be determined by imposing additional properties on the accessibility relations (e.g. transitivity, reflexivity, etc.), which is usually done by stating *axiom schemas*. These characterize different *systems* of modal logic. For now we suffice with a comment on the normal modal logic K, of which all other normal modal logics are extensions. In the system K the axiom schema $K : \Box_i(\alpha \rightarrow \beta) \rightarrow (\Box_i \alpha \rightarrow \Box_i \beta)$ is valid, and the *necessitation* rule $RN : \alpha / \Box_i \alpha$ holds.

For more details on modal logic, we refer the reader to the classic book by Chellas (1980) and the more recent handbook by Blackburn, van Benthem and Wolter (2006).

Preferential Semantics for Modal Logic

Now we present a brief summary of the constructions for preferential reasoning in modal logic as studied by Britz et al. (2011a). (As we shall see in the sequel, the semantics of the defeasible modalities we shall introduce relies heavily on the definitions provided below.)

Definition 3 *Let $\mathcal{M}$ be a class of Kripke models. We define $\mathscr{U}_{\mathcal{M}} := \{(\mathscr{M}, w) \mid \mathscr{M} = \langle W, R, V \rangle \in \mathcal{M}$ and $w \in W\}$. We call $\mathscr{U}_{\mathcal{M}}$ the class of pointed Kripke models from $\mathcal{M}$.*

It is worth noting that pointed Kripke models are not to be viewed as objects, as they are commonly regarded in the context of e.g. description logics (Baader et al. 2007). A set of pointed Kripke models describes the *intention* of a modal statement — cf. Definition 2.

As an example, if $\mathcal{M} = \{\mathscr{M}_1, \mathscr{M}_2\}$, where $\mathscr{M}_1$ and $\mathscr{M}_2$ are as depicted in Figure 1, then $\mathscr{U}_{\mathcal{M}} = \{(\mathscr{M}_1, w), (\mathscr{M}_1, w_2), (\mathscr{M}_1, w_3), (\mathscr{M}_1, w_4), (\mathscr{M}_2, w_1), (\mathscr{M}_2, w_2), (\mathscr{M}_2, w_3), (\mathscr{M}_2, w_4)\}$.

Let S be a set, the elements of which are called *states*, and let $\ell : S \longrightarrow \mathscr{U}_{\mathcal{M}}$ be a *labeling function* mapping every state to a pair $(\mathscr{M}, w)$ where $\mathscr{M} = \langle W, R, V \rangle$ is a Kripke model such that $w \in W$. Moreover, let $\prec \subseteq S \times S$ be a *preference relation* on states, which we assume to be a strict partial order. We say that S satisfies the *smoothness condition* (Kraus, Lehmann, and Magidor 1990) if and only if every subset of S has a $\prec$-minimal element.

Definition 4 (Britz et al. (2011a)) *Let $\mathcal{M}$ be a given class of Kripke models. A preferential model is a triple $\mathscr{P} = \langle S, \ell, \prec \rangle$ where S is a set of states satisfying the smoothness condition, ℓ is a labeling function mapping states to elements of $\mathscr{U}_{\mathcal{M}}$, and $\prec$ is a strict partial order on S, i.e., $\prec$ is irreflexive and transitive.*

Figure 2 shows an example of a (modal) preferential model for $\mathcal{M} = \{\mathscr{M}_1, \mathscr{M}_2\}$, where $\mathscr{M}_1$ and $\mathscr{M}_2$ are as depicted in Figure 1.

$\mathscr{P}$:

$s_7 : (\mathscr{M}_1, w_4),\ s_8 : (\mathscr{M}_2, w_4)$
$s_6 : (\mathscr{M}_2, w_2)$
$s_4 : (\mathscr{M}_1, w_3),\ s_5 : (\mathscr{M}_2, w_3)$
$s_3 : (\mathscr{M}_1, w_2)$
$s_1 : (\mathscr{M}_1, w_1),\ s_2 : (\mathscr{M}_2, w_1)$

most preferred ↓

Figure 2: Example of a preferential model for $\mathcal{M} = \{\mathscr{M}_1, \mathscr{M}_2\}$, where $\mathscr{M}_1$ and $\mathscr{M}_2$ are as depicted in Figure 1.

Given a preferential model $\mathscr{P}$ and $\alpha \in \mathcal{L}$, with $[\![\alpha]\!]$ we denote the set of states satisfying α in $\mathscr{P}$, according to the following definition:

Definition 5 (Satisfaction in Preferential Models) *Let $\mathscr{P} = \langle S, \ell, \prec \rangle$ and $\alpha \in \mathcal{L}$. Then $[\![\alpha]\!] := \{s \in S \mid \ell(s) = (\mathscr{M}, w)$, for some $\mathscr{M} = \langle W, R, V \rangle$ such that $\mathscr{M}, w \Vdash \alpha\}$.*

Given $\alpha \in \mathcal{L}$ and $\mathscr{P}$ a preferential model, we say that α is *satisfiable* in $\mathscr{P}$ if $[\![\alpha]\!] \neq \emptyset$, otherwise α is *unsatisfiable* in $\mathscr{P}$. We say that α is *true* in $\mathscr{P}$ (denoted $\mathscr{P} \Vdash \alpha$) if and only if $[\![\alpha]\!] = S$.

Definition 6 *Let $\mathcal{M}$ be a class of Kripke models and $\mathscr{U}_{\mathcal{M}}$ be the corresponding class of pointed Kripke models from $\mathcal{M}$. The class $\mathcal{M}^{\mathscr{P}}$ of preferential models is the set of all preferential models $\mathscr{P} = \langle S, \ell, \prec \rangle$ such that range$(\ell) \subseteq \mathscr{U}_{\mathcal{M}}$.*

Let $\alpha \in \mathcal{L}$ and let $\mathcal{M}$ be a given class of Kripke models. We say that α is *valid* in $\mathcal{M}^{\mathscr{P}}$ (denoted $\models \alpha$) if and only if α is true in every preferential model $\mathscr{P}$ of $\mathcal{M}^{\mathscr{P}}$, i.e., $\mathscr{P} \Vdash \alpha$ for every $\mathscr{P} \in \mathcal{M}^{\mathscr{P}}$.

Lemma 1 *Let $\alpha \in \mathcal{L}$ (i.e., α is a classical modal formula) and let $\mathcal{M}$ be a class of Kripke models. Then α is valid in $\mathcal{M}$ if and only if α is valid in $\mathcal{M}^{\mathscr{P}}$.*

Definition 7 (Minimality w.r.t. $\prec$) *Let $\mathscr{P} = \langle S, \ell, \prec \rangle$ and let $S' \subseteq S$. With $\min_{\prec} S'$ we denote the minimal elements of S' with respect to $\prec$, i.e., $\min_{\prec} S' = \{x \in S' \mid$ there is no $y \in S'$ such that $y \prec x\}$.*

Given a preferential model $\mathscr{P} = \langle S, \ell, \prec \rangle$, the defeasible statement $\alpha \mathrel{|\!\sim_{\mathscr{P}}} \beta$ holds in $\mathscr{P}$ if and only if every $\prec$-minimal α-state is a β-state, i.e., $\min_{\prec} [\![\alpha]\!] \subseteq [\![\beta]\!]$.

It is worth noting that, in spite of the richer language (modal logic) and the richer semantics (based on possible worlds), defeasible statements here still have the same intuition as mentioned in the Introduction. To witness, the statement $\Diamond \alpha \mathrel{|\!\sim} \Box \beta$ just says that "all normal worlds with an α-successor have only β-successors". That is, any $|\!\sim$-statement still refers only to normality in the premise, or, in this case, of the 'actual' world.

Logics with Defeasible Modalities

Recalling our discussion in the Introduction, we want to be able to state that a given sentence holds in *all* the normal worlds that are accessible, or in *some* of such normal worlds.

This leads us to the definition of 'weaker' versions of modalities, in the sense that both necessity and possibility can now 'fail', or rather have defeasible versions. Through them we shall be able to single out those normal situations that one cannot grasp via standard $\Box$ and $\Diamond$ in the classical case.

We define a more expressive language than $\mathcal{L}$ by extending our modal language with a family of defeasible modal operators $\mathrel{\sim\!\!\Box}_i$ and $\mathrel{\sim\!\!\Diamond}_i$, $1 \leq i \leq n$ (called, respectively, the 'flag' and the 'flame'), where n is the number of classical modalities in the language. The formulas of the extended language are then recursively defined by:

$$\alpha \ ::= p \mid \neg\alpha \mid \alpha \wedge \alpha \mid \Box_i\alpha \mid \mathrel{\sim\!\!\Box}_i\alpha \mid \mathrel{\sim\!\!\Diamond}_i\alpha$$

(As before, the other connectives are defined in terms of $\neg$ and $\wedge$ in the usual way, and $\top$ and $\bot$ are seen as abbreviations. It turns out that each $\mathrel{\sim\!\!\Diamond}_i$ too is the dual of $\mathrel{\sim\!\!\Box}_i$, as we shall see below.) With $\widetilde{\mathcal{L}}$ we denote the set of all formulas of such a richer language.

The semantics of $\widetilde{\mathcal{L}}$ is in terms of our modal preferential models (see Definition 4). As before, given $\alpha \in \widetilde{\mathcal{L}}$ and a preferential model $\mathscr{P}$, with $[\![\alpha]\!]$ we denote the set of all states satisfying α in $\mathscr{P}$.

Definition 8 (Satisfaction Extended) *Let $\mathscr{P} = \langle S, \ell, \prec\rangle$ be a modal preferential model. Then:*

- $[\![\mathrel{\sim\!\!\Box}_i\alpha]\!] := \{s \mid \ell(s) = (\mathscr{M}, w)$, *for* $\mathscr{M} = \langle W, R, V\rangle$, *and* $\min_\prec\{s' \mid \ell(s') = (\mathscr{M}, w')$ *and* $(w, w') \in R_i\} \subseteq [\![\alpha]\!]\}$;
- $[\![\mathrel{\sim\!\!\Diamond}_i\alpha]\!] := \{s \mid \ell(s) = (\mathscr{M}, w)$, *for* $\mathscr{M} = \langle W, R, V\rangle$, *and* $\min_\prec\{s' \mid \ell(s') = (\mathscr{M}, w')$ *and* $(w, w') \in R_i\} \cap [\![\alpha]\!] \neq \emptyset\}$.

The notions of satisfaction in a preferential model, truth (in a model) and validity (in a class of models) are extended to formulas with defeasible modalities in the obvious way.

As alluded to above, we observe that, like in the classical (i.e., non-defeasible) case, the defeasible modal operators $\mathrel{\sim\!\!\Box}$ and $\mathrel{\sim\!\!\Diamond}$ are the dual of each other:

$$\models \mathrel{\sim\!\!\Box}_i\alpha \leftrightarrow \neg\mathrel{\sim\!\!\Diamond}_i\neg\alpha \tag{1}$$

Verification: Let $\mathscr{P} = \langle S, \ell, \prec\rangle$ and let $s \in S$. Then $\ell(s) = (\mathscr{M}, w)$ for some $\mathscr{M} = \langle W, R, V\rangle$ and $w \in W$. $s \in [\![\mathrel{\sim\!\!\Box}_i\alpha]\!]$ if and only if $\min_\prec\{s' \mid \ell(s') = (\mathscr{M}, w')$ and $(w, w') \in R_i\} \subseteq [\![\alpha]\!]$ if and only if $\min_\prec\{s' \mid \ell(s') = (\mathscr{M}, w')$ and $(w, w') \in R_i\} \cap [\![\neg\alpha]\!] = \emptyset$ if and only if $s \notin [\![\mathrel{\sim\!\!\Diamond}_i\neg\alpha]\!]$ if and only if $s \in [\![\neg\mathrel{\sim\!\!\Diamond}_i\neg\alpha]\!]$.

The following property is easy to verify: If there are no most preferred accessible worlds, then there are no accessible worlds at all (and vice versa).

$$\models \mathrel{\sim\!\!\Box}_i\bot \leftrightarrow \Box_i\bot \tag{2}$$

From (2) and contraposition we conclude $\models \Diamond_i\top \leftrightarrow \mathrel{\sim\!\!\Diamond}_i\top$.

The following two equivalences are also worthy of mention (their proofs are straightforward):

$$\models \mathrel{\sim\!\!\Box}_i\top \leftrightarrow \top \text{ and } \models \mathrel{\sim\!\!\Diamond}_i\bot \leftrightarrow \bot \tag{3}$$

The following is the $\mathrel{\sim\!\!\Box}$-version of Axiom Schema *K*.

$$(NK) \ \models \mathrel{\sim\!\!\Box}_i(\alpha \to \beta) \to (\mathrel{\sim\!\!\Box}_i\alpha \to \mathrel{\sim\!\!\Box}_i\beta) \tag{4}$$

Verification: Let $\mathscr{P} = \langle S, \ell, \prec\rangle$ and let $s \in S$. Then $\ell(s) = (\mathscr{M}, w)$ for some $\mathscr{M} = \langle W, R, V\rangle$ and $w \in W$. Let $s \in [\![\mathrel{\sim\!\!\Box}_i(\alpha \to \beta)]\!]$. Then we have (*i*) $\min_\prec\{s' \mid \ell(s') = (\mathscr{M}, w')$ and $(w, w') \in R_i\} \subseteq [\![\alpha \to \beta]\!]$. If $s \in [\![\mathrel{\sim\!\!\Box}_i\alpha]\!]$, then we also have (*ii*) $\min_\prec\{s' \mid \ell(s') = (\mathscr{M}, w')$ and $(w, w') \in R_i\} \subseteq [\![\alpha]\!]$. From (*i*) and (*ii*) it follows that $\min_\prec\{s' \mid \ell(s') = (\mathscr{M}, w')$ and $(w, w') \in R_i\} \subseteq [\![\alpha]\!] \cap [\![\alpha \to \beta]\!] = [\![\alpha \wedge (\alpha \to \beta)]\!] \subseteq [\![\beta]\!]$. Putting the results together gives us $s \in [\![\mathrel{\sim\!\!\Box}_i\beta]\!]$.

The validity below is easy to verify:

$$(NR) \ \models \mathrel{\sim\!\!\Box}_i(\alpha \wedge \beta) \leftrightarrow (\mathrel{\sim\!\!\Box}_i\alpha \wedge \mathrel{\sim\!\!\Box}_i\beta) \tag{5}$$

We also have $\models (\mathrel{\sim\!\!\Box}_i\alpha \vee \mathrel{\sim\!\!\Box}_i\beta) \to \mathrel{\sim\!\!\Box}_i(\alpha \vee \beta)$, but not the other direction of the implication, as can be easily checked.

The following validity is an immediate consequence of our semantics:

$$(N) \ \models \Box_i\alpha \to \mathrel{\sim\!\!\Box}_i\alpha \tag{6}$$

Intuitively, given $i = 1, \ldots, n$, where n is the number of modalities in the language, we want $\Box_i$ and $\mathrel{\sim\!\!\Box}_i$ to be 'tied' together in so far as one is the defeasible (or the 'hard') version of the other.

From duality of $\mathrel{\sim\!\!\Diamond}$ and $\mathrel{\sim\!\!\Box}$ and contraposition of *N* we get:

$$\models \mathrel{\sim\!\!\Diamond}_i\alpha \to \Diamond_i\alpha \tag{7}$$

The following rule of *normal necessitation* (*RNN*) follows from the standard necessitation rule *RN* together with Schema *N* in (6) above:

$$(RNN) \ \frac{\alpha}{\mathrel{\sim\!\!\Box}_i\alpha} \tag{8}$$

From satisfaction of (1), (4) and (5), one can see that the logic of our defeasible modalities shares properties commonly characterizing the so-called *normal* modal logics (Chellas 1980). In particular, we have that the following rule holds:

$$(NRK) \ \frac{(\alpha_1 \wedge \ldots \wedge \alpha_k) \to \beta}{(\mathrel{\sim\!\!\Box}_i\alpha_1 \wedge \ldots \wedge \mathrel{\sim\!\!\Box}_i\alpha_k) \to \mathrel{\sim\!\!\Box}_i\beta} \ (k \geq 0) \tag{9}$$

Verification: Assume that $\models (\alpha_1 \wedge \ldots \wedge \alpha_k) \to \beta$. Then by Rule *RNN* we conclude $\models \mathrel{\sim\!\!\Box}_i((\alpha_1 \wedge \ldots \wedge \alpha_k) \to \beta)$. This and Schema *NK* give us $\models \mathrel{\sim\!\!\Box}_i(\alpha_1 \wedge \ldots \wedge \alpha_k) \to \mathrel{\sim\!\!\Box}_i\beta$. Iterated applications of *NR* tell us that the antecedent of the latter is logically equivalent to $\mathrel{\sim\!\!\Box}_i\alpha_1 \wedge \ldots \wedge \mathrel{\sim\!\!\Box}_i\alpha_k$. From this result and substitution of equivalents we conclude that $\models (\mathrel{\sim\!\!\Box}_i\alpha_1 \wedge \ldots \wedge \mathrel{\sim\!\!\Box}_i\alpha_k) \to \mathrel{\sim\!\!\Box}_i\beta$.

The observant reader would have noticed that we assume we have as many defeasible modalities as we have classical ones. That is, for each $\Box_i$, a corresponding $\mathrel{\sim\!\!\Box}_i$ (its defeasible version) is assumed. Moreover they are both linked together via Schema *N* in (6). In principle, from a technical point of view, nothing precludes us from having defeasible modalities with no corresponding classical version or the other way round. The latter case is easily dealt with by simply not having $\mathrel{\sim\!\!\Box}_i$ for some i for which $\Box_i$ is present in the language. The former case, on the other hand, would require a slight

elaboration of the semantics as, currently, satisfiability of $\widetilde{\Box}$-formulas call upon the accessibility relation R_i, usually associated with a $\Box_i$-modality. Even though one can make a case for only wanting the defeasible version of a given modality to be available, it deviates from our stated aim of having defeasible *versions* of the (already existing) modalities in our language and therefore we do not investigate this further here.

From the perspective of knowledge representation and reasoning, it becomes important to address the question of what it means for an $\widetilde{\mathcal{L}}$-sentence to be *entailed* from an $\widetilde{\mathcal{L}}$-knowledge base.

An $\widetilde{\mathcal{L}}$-*knowledge base* is a (possibly infinite) set of sentences $\mathcal{K} \subseteq \widetilde{\mathcal{L}}$. Given a modal preferential model $\mathscr{P}$, we extend the notion of satisfaction to knowledge bases in the obvious way: $\mathscr{P} \Vdash \mathcal{K}$ if and only if $\mathscr{P} \Vdash \alpha$ for every $\alpha \in \mathcal{K}$.

Definition 9 (Preferential Modal Entailment) *Let $\mathcal{K} \subseteq \widetilde{\mathcal{L}}$ and let $\alpha \in \widetilde{\mathcal{L}}$. We say that $\mathcal{K}$ entails α in the class $\mathcal{M}^{\mathscr{P}}$ of preferential models (denoted $\mathcal{K} \models \alpha$) if and only if for every $\mathscr{P} \in \mathcal{M}^{\mathscr{P}}$, if $\mathscr{P} \Vdash \mathcal{K}$, then $\mathscr{P} \Vdash \alpha$.*

Given this notion of entailment, its associated *consequence relation* is defined as:

$$Cn(\mathcal{K}) \equiv_{\text{def}} \{\alpha \mid \mathcal{K} \models \alpha\} \tag{10}$$

It is easy to see that the consequence relation $Cn(\cdot)$ as defined in (10) above is a Tarskian consequence relation:

Theorem 1 *Let $Cn(\cdot)$ be a consequence relation defined in terms of preferential modal entailment. Then $Cn(\cdot)$ satisfies the following properties:*

- $\mathcal{K} \subseteq Cn(\mathcal{K})$ *(Inclusion)*
- $Cn(\mathcal{K}) = Cn(Cn(\mathcal{K}))$ *(Idempotency)*
- *If $\mathcal{K}_1 \subseteq \mathcal{K}_2$, then $Cn(\mathcal{K}_1) \subseteq Cn(\mathcal{K}_2)$* *(Monotonicity)*

That is, in spite of the defeasibility features of $\widetilde{\Box}$, we end up with a logic that is monotonic, just as in Kraus et al.'s preferential entailment (Kraus, Lehmann, and Magidor 1990).

Below is an interesting result relating truth of $\widetilde{\mathcal{L}}$-sentences in a preferential model with the defeasible consequence relation induced by the model (cf. paragraph after Definition 7).

Lemma 2 *Let $\alpha \in \widetilde{\mathcal{L}}$ and $\mathscr{P}$ be a preferential model. Then $\mathscr{P} \Vdash \alpha$ if and only if $\neg\alpha \mathrel{|\!\!\sim}_{\mathscr{P}} \bot$.*

This result raises the obvious question on whether we can reduce entailment of $\widetilde{\mathcal{L}}$-sentences to that of $\mathrel{|\!\!\sim}$-statements. To make this more precise, we need some definitions. We say that a preferential model $\mathscr{P}$ satisfies a defeasible statement $\alpha \mathrel{|\!\!\sim} \beta$ if and only if $\alpha \mathrel{|\!\!\sim}_{\mathscr{P}} \beta$ holds. $\mathscr{P}$ satisfies a set of such defeasible statements if $\mathscr{P}$ satisfies each of them. Given a set X of defeasible statements, we say that X (preferentially) entails $\alpha \mathrel{|\!\!\sim} \beta$ (denoted $X \models \alpha \mathrel{|\!\!\sim} \beta$) if every preferential model satisfying all the statements in X also satisfies $\alpha \mathrel{|\!\!\sim} \beta$. (Given Lemma 2 it is not hard to see that $\models$ here is exactly the same entailment relation from Definition 9, just restated in terms of $\mathrel{|\!\!\sim}$-statements.)

Definition 10 *Let $\mathcal{K} \subseteq \widetilde{\mathcal{L}}$. $\mathcal{K}^{\mathrel{|\!\sim}} := \{\neg\alpha \mathrel{|\!\!\sim} \bot \mid \alpha \in \mathcal{K}\}$.*

Theorem 2 *$\mathcal{K} \models \alpha$ if and only if $\mathcal{K}^{\mathrel{|\!\sim}} \models \neg\alpha \mathrel{|\!\!\sim} \bot$.*

Hence, preferential entailment in $\widetilde{\mathcal{L}}$ reduces to preferential entailment of $\mathrel{|\!\!\sim}$-statements in the language of $\widetilde{\mathcal{L}}$. An immediate consequence of this is that the existence of a sound and complete KLM-style $\mathrel{|\!\!\sim}$-based proof system (Kraus, Lehmann, and Magidor 1990) for $\widetilde{\mathcal{L}}$ would define a decision procedure for $\widetilde{\mathcal{L}}$. At present we can only conjecture that such a proof system exists.

Examples of Defeasible Modes of Inference

Let us assume the following scenario depicting a nuclear power station (Britz, Meyer, and Varzinczak 2011a): In a particular power plant there is an atomic pile and a cooling system, both of which can be either on or off. An agent is in charge of detecting hazardous situations and preventing the plant from malfunctioning (Figure 3).

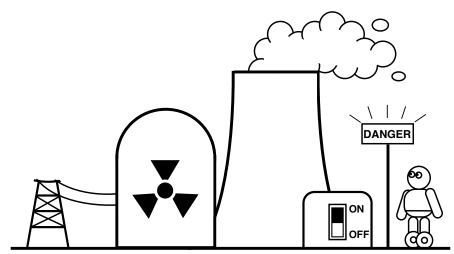

Figure 3: The nuclear power station and its controlling agent.

In what follows we shall illustrate our constructions from previous sections in reasoning about action and in epistemic reasoning using the aforementioned scenario.

Dynamic Defeasibility

We find in the AI literature a fair number of modal-based formalisms for reasoning about actions and change (De Giacomo and Lenzerini 1995; Zhang and Foo 2001; Castilho, Herzig, and Varzinczak 2002; Demolombe, Herzig, and Varzinczak 2003; Herzig and Varzinczak 2007). These are essentially variants of the modal logic K we presented in the Preliminaries Section. Modal operators are determined by a (finite) set of *actions* $\mathcal{A} = \{a_1, \ldots, a_n\}$: For each $a \in \mathcal{A}$, there is associated a modal operator $\Box_a$. Given a Kripke model, $R_a \subseteq W \times W$ is therefore meant to represent possible executions of an (ontic) action a at specific worlds $w \in W$, i.e., R_a is the specification of a's behavior in a transition system. Hence whenever $(w, w') \in R_a$, w' is a *possible outcome* of doing a in w. Formulas of the form $\Box_a\alpha$ are used to specify the *effects* of actions and they are read "after *every* execution of action a, the formula α holds". The operator $\Diamond_a$ is mostly used to specify the *executability* of actions: $\Diamond_a\top$ reads "there is a possible execution of action a".

In our nuclear power plant example, let $\mathcal{P} = \{\mathsf{p}, \mathsf{c}, \mathsf{h}\}$ be a set of propositions, where p stands for "the atomic pile is on", c for "the cooling system is on", and h for "hazardous

situation". Moreover, let $\mathcal{A} = \{\mathsf{f}, \mathsf{m}\}$ be a set of atomic actions, where f stands for "flipping the pile switch", and m for "malfunction".

Assume that we are given $\mathcal{M} = \{\mathscr{M}_1, \mathscr{M}_2\}$, where $\mathscr{M}_1$ and $\mathscr{M}_2$ are as depicted in Figure 4.

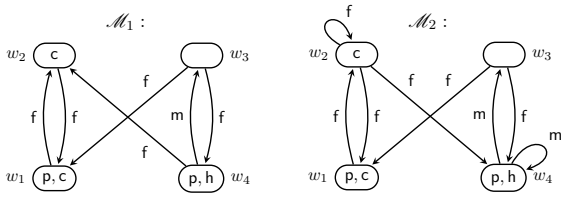

Figure 4: Kripke models depicting the behavior of actions in our nuclear power station scenario.

Hence $\mathscr{U}_{\mathcal{M}} = \{(\mathscr{M}_i, w_j) \mid i \in \{1,2\}, j \in \{1,2,3,4\}\}$. We construct a preferential model (Definition 4) in which to check the satisfiability and truth of a few sentences. The purpose is to illustrate the semantics of our notion of defeasibility in an action context rather than to present a comprehensive modeling of the nuclear power plant scenario.

Assume $S = \{s_i \mid 1 \leq i \leq 8\}$, and let a labeling function ℓ be such that $\ell(s_1) = (\mathscr{M}_1, w_1)$, $\ell(s_2) = (\mathscr{M}_2, w_1)$, $\ell(s_3) = (\mathscr{M}_1, w_2)$, $\ell(s_4) = (\mathscr{M}_1, w_3)$, $\ell(s_5) = (\mathscr{M}_2, w_3)$, $\ell(s_6) = (\mathscr{M}_2, w_2)$, $\ell(s_7) = (\mathscr{M}_1, w_4)$, and $\ell(s_8) = (\mathscr{M}_2, w_4)$. The order $\prec$ is given by: $s_1 \prec s_3$, $s_2 \prec s_3$, $s_3 \prec s_4$, $s_3 \prec s_5$, $s_4 \prec s_6$, $s_5 \prec s_6$, $s_6 \prec s_7$, and $s_6 \prec s_8$. Figure 5 below depicts the preferential model $\mathscr{P} = \langle S, \ell, \prec \rangle$.

$$\mathscr{P}: \quad \begin{array}{|c|} \hline s_7 : (\mathscr{M}_1, w_4),\ s_8 : (\mathscr{M}_2, w_4) \\ \hline s_6 : (\mathscr{M}_2, w_2) \\ \hline s_4 : (\mathscr{M}_1, w_3),\ s_5 : (\mathscr{M}_2, w_3) \\ \hline s_3 : (\mathscr{M}_1, w_2) \\ \hline s_1 : (\mathscr{M}_1, w_1),\ s_2 : (\mathscr{M}_2, w_1) \\ \hline \end{array} \quad \textit{most preferred} \downarrow$$

Figure 5: Preferential model for the power plant scenario.

The rationale of this partial order is as follows: The utility company selling the electricity generated by the power plant tries as far as possible to keep both the pile and the cooling system on, ensuring that the pile can be easily switched off (states s_1 and s_2); sometimes the company has to switch the pile off for maintenance but then tries to keep the cooler running, preferably if turning the pile on again does not cause a fault in the cooling system (state s_3); more rarely the company needs to switch off both the pile and the cooler, e.g. when the latter needs maintenance (states s_4 and s_5); in an exceptional situation, turning the pile on not only may not produce its intended effect but can also interfere with the cooler switching it off (state s_6); and, finally, only in very exceptional situations would the pile be on while the cooler is off, e.g. during a serious malfunction (states s_7 and s_8).

In the preferential model $\mathscr{P}$ depicted above, one can check that $s_6 \in [\![\mathsf{h} \wedge \mathring{\Diamond}_{\mathsf{f}} \neg \mathsf{h}]\!]$: at s_6 we have a hazardous situation, but it is possible to switch the pile off having as a normal effect a safe condition. We have that s_7 does not satisfy $\mathrel{\sim}_{\mathsf{m}} \bot$, which is satisfied by s_1: at s_1 a malfunction cannot occur (cf. Axioms 2 and 6). In $\mathscr{P}$ we have $\mathscr{P} \Vdash \neg \mathsf{p} \rightarrow \mathrel{\sim}_{\mathsf{f}} \mathsf{p}$ (the normal outcome of switching the pile on is it being on), but $\mathscr{P} \nVdash \neg \mathsf{p} \rightarrow \Box_{\mathsf{f}} \mathsf{p}$ (see state s_6). We also have $\mathscr{P} \Vdash \mathsf{c} \rightarrow \mathrel{\sim}_{\mathsf{f}} \neg \mathsf{h}$ (if the cooler is on, the normal result of switching the pile is a safe situation), but $\mathscr{P} \nVdash \mathsf{c} \rightarrow \Box_{\mathsf{f}} \neg \mathsf{h}$. Finally we also have $\mathscr{P} \Vdash \mathsf{h} \rightarrow \mathring{\Diamond}_{\mathsf{m}} \top$: In any hazardous situation a meltdown is likely to happen.

So far we have illustrated the preferential semantics of $\widetilde{\mathcal{L}}$-statements using specific Kripke models and preference orders. In a knowledge representation context, though, we are interested in preferential entailment from an $\widetilde{\mathcal{L}}$-theory or knowledge base. The latter determines the preferential models that are permissible from the standpoint of the knowledge engineer. To illustrate this, consider the following $\widetilde{\mathcal{L}}$-knowledge base:

$$\mathcal{K} = \left\{ \begin{array}{c} (\mathsf{p} \wedge \neg \mathsf{c}) \leftrightarrow \mathsf{h},\ \mathsf{h} \rightarrow \mathring{\Diamond}_{\mathsf{m}} \top, \\ \mathsf{p} \rightarrow \mathrel{\sim}_{\mathsf{f}} \neg \mathsf{p},\ \mathsf{c} \rightarrow \mathrel{\sim}_{\mathsf{f}} \mathsf{c},\ \mathring{\Diamond}_{\mathsf{f}} \top \end{array} \right\}$$

$\mathcal{K}$ basically says that "a hazardous situation is one in which the pile is on and the cooler off", "a hazardous situation may normally lead to a malfunction", "if the pile is on, then flipping its switch normally switches it off", "if the cooler is on, then switching the pile normally does not affect it" and "one can normally flip the pile switch". (It is not hard to check that all the formulas in $\mathcal{K}$ are true in the preferential model $\mathscr{P}$ of Figure 5 above.) We can then conclude $\mathcal{K} \models \mathsf{p} \rightarrow \mathrel{\sim}_{\mathsf{f}} \neg \mathsf{h}$, $\mathcal{K} \models \mathrel{\sim}_{\mathsf{m}} \bot \rightarrow (\neg \mathsf{p} \vee \mathsf{c})$ and $\mathcal{K} \models (\mathsf{p} \vee \mathsf{c}) \rightarrow \mathrel{\sim}_{\mathsf{f}} \neg \mathsf{h}$, using the sound $\widetilde{\mathcal{L}}$-inference rules and validities presented in the previous section.

Note that, in the knowledge base $\mathcal{K}$ above, the formula $\mathsf{p} \rightarrow \mathrel{\sim}_{\mathsf{f}} \neg \mathsf{p}$ says something different than what the statement $\mathsf{p} \mathrel{|\!\sim} \Box_{\mathsf{f}} \neg \mathsf{p}$ would convey in a $\mathrel{|\!\sim}$-based formalization of this scenario as we investigated in previous work (Britz, Meyer, and Varzinczak 2011a). The latter says that "in a *normal* situation where the pile is on, *every* outcome of the flipping action switches the pile off", whereas the former specifies that "in *any* situation, the *normal* effect of flipping is the pile being off".

Epistemic Defeasibility

Another family of logics that are of great interest from the standpoint of AI is that of epistemic logics, which allow for reasoning about knowledge (Fagin et al. 1995).

The language of (modal) epistemic logic contains a (finite) set of *agents* $\mathcal{A} = \{A_1, \ldots, A_n\}$. For each agent $A \in \mathcal{A}$ there is a *knowledge operator* $\Box_A$. Given a Kripke model, $R_A \subseteq W \times W$ represents all epistemic possibilities from agent A's standpoint. A formula of the form $\Box_A \alpha$ is therefore used to specify A's *knowledge* about the world and it is read as "agent A *knows* that α is the case".

The core of epistemic logic is the normal multi-modal logic K_m. Hence, the following version of axiom schema *K* is valid: $\Box_A \alpha \wedge \Box_A (\alpha \rightarrow \beta) \rightarrow \Box_A \beta$, i.e., "if A knows both α and $\alpha \rightarrow \beta$, then she also knows β". Stronger epistemic logics are obtained by adding additional schemata, expressing specific desired properties of knowledge, to the basic

system K. Since K is at the heart of these logics, we shall suffice with it in our exposition below.

In what follows we turn our attention to the application of defeasible modalities in epistemic reasoning. In this context, given an agent A, we shall read a formula of the form $\mathrel{\sim\!\!\square}_A\alpha$ as "A knows that normally α".

Still in our power plant scenario, let us assume that we have two agents, say A and B. The set $\mathcal{P}$ is as in the previous section, with the propositions p, c and h keeping their previous intuition.

Here we do a similar exercise to that of the previous section. To that matter, assume that we are given a class of models $\mathcal{M} = \{\mathscr{M}_1, \mathscr{M}_2\}$, where $\mathscr{M}_1$ and $\mathscr{M}_2$ are now as depicted in Figure 6.

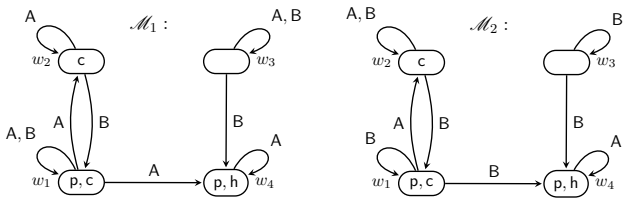

Figure 6: Kripke models depicting knowledge of two agents in our nuclear power station scenario.

Assume again that $S = \{s_i \mid 1 \leq i \leq 8\}$, and let a labeling function ℓ be such that $\ell(s_1) = (\mathscr{M}_1, w_1)$, $\ell(s_2) = (\mathscr{M}_2, w_1)$, $\ell(s_3) = (\mathscr{M}_1, w_2)$, $\ell(s_4) = (\mathscr{M}_2, w_2)$, $\ell(s_5) = (\mathscr{M}_1, w_3)$, $\ell(s_6) = (\mathscr{M}_2, w_3)$, $\ell(s_7) = (\mathscr{M}_1, w_4)$, and $\ell(s_8) = (\mathscr{M}_2, w_4)$. The order $\prec$ is given by: $s_i \prec s_{i+2}$ and $s_i \prec s_{i+3}$, for $i \in \{1, 3, 5\}$, and $s_j \prec s_{j+1}$ and $s_j \prec s_{j+2}$, for $j \in \{2, 4, 6\}$. Figure 7 below depicts the preferential model $\mathscr{P} = \langle S, \ell, \prec\rangle$ generated like this.

$$\mathscr{P}: \quad \begin{array}{|c|}\hline s_7 : (\mathscr{M}_1, w_4),\ s_8 : (\mathscr{M}_2, w_4) \\ \hline s_5 : (\mathscr{M}_1, w_3),\ s_6 : (\mathscr{M}_2, w_3) \\ \hline s_3 : (\mathscr{M}_1, w_2),\ s_4 : (\mathscr{M}_2, w_2) \\ \hline s_1 : (\mathscr{M}_1, w_1),\ s_2 : (\mathscr{M}_2, w_1) \\ \hline \end{array} \quad \textit{most preferred} \downarrow$$

Figure 7: A preferential model for knowledge in the power plant scenario.

The rationale behind such a partial order is quite similar to that of the previous section. Here states in which the pile and the cooling system are on and in which at least one agent is aware of what is going on there are the most preferred (states s_1 and s_2). These are followed by those in which the pile is off while the cooler is still running and agent A knows that the pile is off (states s_3 and s_4). States s_5 and s_6 capture the plant being in maintenance mode, while s_7 and s_8 are the least normal situations, namely when a meltdown is imminent and not all the agents know that is the case.

In this preferential model $\mathscr{P}$, one can verify that $s_1 \in [\![\mathrel{\sim\!\!\square}_\mathsf{A}\mathsf{c} \wedge \mathrel{\sim\!\!\square}_\mathsf{A}\mathrel{\sim\!\!\square}_\mathsf{A}\mathsf{c}]\!]$, i.e., agent A knows that normally the cooling system is on and, moreover, knows that normally it knows this. We have that $s_2 \in [\![\mathrel{\sim\!\!\diamond}_\mathsf{B}\mathrel{\sim\!\!\square}_\mathsf{A}\neg\mathsf{h}]\!]$: in s_2 B conceives that normally A knows that normally it is a safe situation. In $\mathscr{P}$ we have $\mathscr{P} \Vdash \mathsf{p} \wedge \neg\mathsf{c} \rightarrow \mathrel{\sim\!\!\square}_\mathsf{A}\mathsf{h}$ (if the pile is on but the cooler is off, then agent A knows that normally it is a hazardous situation). We also have $\mathscr{P} \Vdash \mathsf{p} \wedge \mathsf{c} \rightarrow \mathrel{\sim\!\!\square}_\mathsf{A}\neg\mathsf{h}$, but $\mathscr{P} \not\Vdash \mathsf{p} \wedge \mathsf{c} \rightarrow \square_\mathsf{A}\neg\mathsf{h}$. It can also be checked that $\mathscr{P} \Vdash \mathrel{\sim\!\!\square}_\mathsf{A}\mathsf{p} \rightarrow \square_\mathsf{B}\mathrel{\sim\!\!\square}_\mathsf{A}\mathsf{p}$ (if A knows that normally the pile is on, then B knows that this is the case). As expected, $\mathscr{P} \not\Vdash \mathrel{\sim\!\!\square}_\mathsf{A}\mathsf{p} \rightarrow \square_\mathsf{B}\square_\mathsf{A}\mathsf{p}$.

We end this section with another illustration of entailment from an $\widetilde{\mathcal{L}}$-theory. Consider the following knowledge base (for illustrative purposes we suffice with a formalization of agent B's knowledge only):

$$\mathcal{K} = \left\{ \begin{array}{c} (\mathsf{p} \wedge \neg\mathsf{c}) \rightarrow \mathsf{h},\ \mathsf{h} \rightarrow \square_\mathsf{B}\mathsf{h}, \\ \mathrel{\sim\!\!\square}_\mathsf{B}(\mathsf{p} \rightarrow \mathsf{c}),\ (\mathsf{p} \wedge \mathsf{c}) \rightarrow \mathrel{\sim\!\!\square}_\mathsf{B}\neg\mathsf{h} \end{array} \right\}$$

The intuition behind $\mathcal{K}$ here is that "if the pile is on and the cooler off, then we have a hazardous situation", "in a hazardous situation, agent B is aware of it for sure", "agent B knows that normally if the pile is on then so is the cooling system", and "if the pile and the cooler are both on, then B knows that normally it is not a hazardous situation". (It can be easily checked that the preferential model from Figure 7 above is a model of $\mathcal{K}$.) Given this knowledge base, the following holds: $\mathcal{K} \models (\mathsf{p} \wedge \neg\mathsf{c}) \rightarrow \mathrel{\sim\!\!\square}_\mathsf{B}\mathsf{h}$ and $\mathcal{K} \models \square_\mathsf{B}\mathsf{p} \rightarrow \mathrel{\sim\!\!\square}_\mathsf{B}\mathrel{\sim\!\!\square}_\mathsf{B}\neg\mathsf{h}$.

Discussion and Related Work

There is a substantial body of work on defeasible knowledge, actions, obligations, beliefs, *et cetera*. What distinguishes our work is that we modify the meaning of modalities through the introduction of a preferential semantics. This semantics allows us to introduce new, defeasible variants of existing modalities.

We have not entered here into a philosophical debate on the nature of knowledge, such as whether knowledge admits defeasibility. We have also not yet investigated the relevance of our defeasible modalities in modeling notions such as justifications (Artemov 2008) or *prima facie* obligations (Asher and Bonevac 1996; Nute 1997). Our focus in this paper has been on the formal semantics of defeasible modalities, and we have explored their usefulness in modeling dynamic and epistemic defeasibility, but their employment in broader range of philosophical application areas such deontic, doxastic and justification logics can be explored in much greater depth.

Baltag and Smets (2008) employ preference orders to define multi-agent epistemic and doxastic *plausibility frames*. These are essentially a special case of Kripke frames, with each accessibility relation induced by a corresponding preference order and linked to an agent. This results in modalities of knowledge, (conditional) belief and safe belief that are closely related to our defeasible modalities.

There are, however, three essential differences between their proposal and ours: (*i*) Their work is presented within the context of dynamic epistemic logic (van Ditmarsch, van der Hoek, and Kooi 2007), and the semantics of their epistemic and doxastic modalities are developed with this specific context in mind; (*ii*) their plausibility orders are subjective orders linked to agents, that determine the agents'

knowledge and beliefs, and (*iii*) an agent's beliefs are determined by what the agent deems epistemically possible. Minimality, or *doxastic appearance*, is therefore determined relative to an epistemic context, which is induced as an equivalence relation on states.

Our work differs from Baltag and Smets' in that (*i*) it offers a preferential semantic framework independent of a specific application area, (*ii*) we assume a single preference order which forms an integral part of the underlying semantics of the background ontology, and (*iii*) the preference order informs the meaning of existing modalities by considering minimality in accessible worlds, where accessibility is determined independently from the preference order.

As we have seen, Britz et al. also propose a general semantic framework for preferential modal logics, but they focus on defeasible arguments rather than on defeasible modalities. As such, the semantics introduced there provides a foundation for the semantics of defeasible modalities, but the syntax of preferential modal logic does not suffice to define preferential modalities such as ours.

In a recent investigation focussing only on rational consequence relations, Booth et al. (2012) introduce an operator with which one can refer directly in the language to those most typical situations in which a given sentence is true. For instance, in their enriched language, a sentence of the form $\overline{\alpha}$ refers to the 'most typical' α-worlds in a semantics similar to ours. One of the advantages of such an extension is the possibility to make statements of the kind "all normal α-worlds are normal β-worlds", thereby shifting the focus of normality from the antecedent by also allowing us to talk about normality in the consequent. This additional expressivity can also be obtained by the addition of the modality $\Box$ of Modular Gödel-Löb logic to express normality syntactically (Giordano et al. 2005; Britz, Heidema, and Labuschagne 2009):

$$\overline{\alpha} \equiv_{\text{def}} \Box\neg\alpha \wedge \alpha \tag{11}$$

Despite the gain in expressivity, both these proposals remain propositional in nature in that the only modality allowed is the one with semantics determined by the preference order. Britz et al. (2011a) extended propositional preferential reasoning to the modal case, but the modalities under consideration there remain classical — their meaning remains as in propositional modal logic, despite the underlying preferential semantics of the logic due to the extension of the language with conditional statements of the form $\alpha \mathrel{|\!\sim} \beta$. Here we have followed an alternative route by investigating additional, defeasible modalities, which can be added to a given modal language by adopting a preferential semantics.

Concluding Remarks

The defeasible modalities introduced in this paper refer to the relative normality of accessible worlds, unlike syntactic characterizations of normality such as those discussed above (Crocco and Lamarre 1992; Boutilier 1994; Giordano et al. 2005; Baltag and Smets 2008; Giordano et al. 2009; Booth, Meyer, and Varzinczak 2012), which refer to the relative normality of worlds in which a given sentence is true, or $\mathrel{|\!\sim}$, which refers to the relative normality of the worlds in which the premise is true.

We have seen that the modal logics obtained through the addition of $\boxbslash_i$ are monotonic, but can be extended to include a non-monotonic conditional $\mathrel{|\!\sim}$. However, such extensions do not make the addition of $\boxbslash_i$ a superfluous extension to the language, since $\boxbslash_i$ cannot be expressed in terms of $\mathrel{|\!\sim}$.

One avenue for future research is therefore integrating $\boxbslash_i$ with modal preferential reasoning, since this would allow for the expression of both defeasible arguments and defeasible modalities. More pressing, though, is the need for a decision procedure for $\widetilde{\mathcal{L}}$. Once this is in place, a deeper exploration of applications in various modal logics is warranted.

Finally, from a knowledge representation and reasoning perspective, when one deals with knowledge bases, issues related to modularization (Herzig and Varzinczak 2004; 2005a; 2005b; 2006), knowledge base update and repair (Herzig, Perrussel, and Varzinczak 2006; Varzinczak 2008; 2010) as well as knowledge base maintenance and versioning (Franconi, Meyer, and Varzinczak 2010) become important. These are threads worthy of investigation in the richer framework of defeasible modalities.

Acknowledgements

This work was partially funded by Project number 247601, Net2: Network for Enabling Networked Knowledge, from the FP7-PEOPLE-2009-IRSES call.

The work of Ivan Varzinczak was supported by the National Research Foundation under Grant number 81225.

References

Artemov, S. 2008. The logic of justification. *The Review of Symbolic Logic* 1(4):477–513.

Asher, N., and Bonevac, D. 1996. Prima facie obligation. *Studia Logica* 57(1):19–45.

Baader, F.; Calvanese, D.; McGuinness, D.; Nardi, D.; and Patel-Schneider, P., eds. 2007. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2 edition.

Baltag, A., and Smets, S. 2008. A qualitative theory of dynamic interactive belief revision. In Bonanno, G.; van der Hoek, W.; and Wooldridge, M., eds., *Logic and the Foundations of Game and Decision Theory (LOFT7)*, number 3 in Texts in Logic and Games, 13–60. Amsterdam University Press.

Blackburn, P.; van Benthem, J.; and Wolter, F. 2006. *Handbook of Modal Logic*. Elsevier North-Holland.

Booth, R.; Meyer, T.; and Varzinczak, I. 2012. PTL: A propositional typicality logic. Submitted.

Boutilier, C. 1994. Conditional logics of normality: A modal approach. *Artificial Intelligence* 68(1):87–154.

Britz, K.; Heidema, J.; and Labuschagne, W. 2009. Semantics for dual preferential entailment. *Journal of Philosophical Logic* 38:433–446.

Britz, K.; Heidema, J.; and Meyer, T. 2008. Semantic preferential subsumption. In Lang, J., and Brewka, G., eds.,

Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR), 476–484. AAAI Press/MIT Press.

Britz, K.; Meyer, T.; and Varzinczak, I. 2011a. Preferential reasoning for modal logics. *Electronic Notes in Theoretical Computer Science* 278:55–69. Proceedings of the 7th Workshop on Methods for Modalities (M4M'2011).

Britz, K.; Meyer, T.; and Varzinczak, I. 2011b. Semantic foundation for preferential description logics. In Wang, D., and Reynolds, M., eds., *Proceedings of the 24th Australasian Joint Conference on Artificial Intelligence*, number 7106 in LNAI, 491–500. Springer.

Castilho, M.; Herzig, A.; and Varzinczak, I. 2002. It depends on the context! A decidable logic of actions and plans based on a ternary dependence relation. In *Proceedings of the 9th International Workshop on Nonmonotonic Reasoning (NMR)*.

Chellas, B. 1980. *Modal logic: An introduction*. Cambridge University Press.

Crocco, G., and Lamarre, P. 1992. On the connections between nonmonotonic inference systems and conditional logics. In Nebel, R.; Rich, C.; and Swartout, W., eds., *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 565–571. Morgan Kaufmann Publishers.

De Giacomo, G., and Lenzerini, M. 1995. PDL-based framework for reasoning about actions. In Gori, M., and Soda, G., eds., *Proceedings of the 4th Congress of the Italian Association for Artificial Intelligence (IA*AI)*, number 992 in LNAI, 103–114. Springer-Verlag.

Demolombe, R.; Herzig, A.; and Varzinczak, I. 2003. Regression in modal logic. *Journal of Applied Non-Classical Logics* 13(2):165–185.

Fagin, R.; Halpern, J.; Moses, Y.; and Vardi, M. 1995. *Reasoning about Knowledge*. MIT Press.

Franconi, E.; Meyer, T.; and Varzinczak, I. 2010. Semantic diff as the basis for knowledge base versioning. In *Proceedings of the 13th International Workshop on Nonmonotonic Reasoning (NMR)*.

Gettier, E. 1963. Is justified true belief knowledge? *Analysis* 23(6):121–123.

Giordano, L.; Gliozzi, V.; Olivetti, N.; and Pozzato, G. 2005. Analytic tableaux for KLM preferential and cumulative logics. In Sutcliffe, G., and Voronkov, A., eds., *Proceedings of LPAR 2005*, volume 3835 of *LNAI*. Berlin, Heidelberg: Springer-Verlag. 666–681.

Giordano, L.; Olivetti, N.; Gliozzi, V.; and Pozzato, G. 2009. $\mathcal{ALC} + T$: a preferential extension of description logics. *Fundamenta Informaticae* 96(3):341–372.

Herzig, A., and Varzinczak, I. 2004. Domain descriptions should be modular. In López de Mántaras, R., and Saitta, L., eds., *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, 348–352. IOS Press.

Herzig, A., and Varzinczak, I. 2005a. Cohesion, coupling and the meta-theory of actions. In Kaelbling, L., and Saffiotti, A., eds., *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, 442–447. Morgan Kaufmann Publishers.

Herzig, A., and Varzinczak, I. 2005b. On the modularity of theories. In Schmidt, R.; Pratt-Hartmann, I.; Reynolds, M.; and Wansing, H., eds., *Advances in Modal Logic*, 5, 93–109. King's College Publications.

Herzig, A., and Varzinczak, I. 2006. A modularity approach for a fragment of $\mathcal{ALC}$. In Fisher, M.; van der Hoek, W.; Konev, B.; and Lisitsa, A., eds., *Proceedings of the 10th European Conference on Logics in Artificial Intelligence (JELIA)*, number 4160 in LNAI, 216–228. Springer-Verlag.

Herzig, A., and Varzinczak, I. 2007. Metatheory of actions: beyond consistency. *Artificial Intelligence* 171:951–984.

Herzig, A.; Perrussel, L.; and Varzinczak, I. 2006. Elaborating domain descriptions. In Brewka, G.; Coradeschi, S.; Perini, A.; and Traverso, P., eds., *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI)*, 397–401. IOS Press.

Kracht, M., and Wolter, F. 1991. Properties of independently axiomatizable bimodal logics. *Journal of Symbolic Logic* 56(4):1469–1485.

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

Lehmann, D., and Magidor, M. 1992. What does a conditional knowledge base entail? *Artificial Intelligence* 55:1–60.

Lehrer, K., and Paxson, T. 1969. Undefeated justified true belief. *The Journal of Philosophy* 66(8):225–237.

Nute, D., ed. 1997. *Defeasible Deontic Logic*, volume 263 of *Synthèse Library*. Kluwer Academic Publishers.

van Ditmarsch, H.; van der Hoek, W.; and Kooi, B. 2007. *Dynamic Epistemic Logic*. Springer.

Varzinczak, I. 2008. Action theory contraction and minimal change. In Lang, J., and Brewka, G., eds., *Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 651–661. AAAI Press/MIT Press.

Varzinczak, I. 2010. On action theory change. *Journal of Artificial Intelligence Research* 37:189–246.

Zhang, D., and Foo, N. 2001. EPDL: A logic for causal reasoning. In Nebel, B., ed., *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, 131–138. Morgan Kaufmann Publishers.