Safety Logics II: Normative Safety

John Bell and Zhisheng Huang [1]

Abstract. In this paper we continue our analysis and formalisation of reasoning about safety. In a companion paper we distinguish between *absolute* safety and *normative* safety, and use Dynamic Logic to develop a formal possible-worlds semantics and a logic for absolute safety. In this paper, we begin by outlining this theory. We then introduce *Defeasible Dynamic Logic* in order to give possible-worlds semantics and a logic for normative safety. We then define a preferential entailment relation defined in order to represent commonsense reasoning about the normal termination of actions. We conclude with a discussion of the relationship between safety, obligation, rationality and risk, and outline some extensions to the present work.

1 Introduction

Attempting to develop a formal theory of reasoning about safety is a complex task; as reasoning of this kind may involve commonsense reasoning about action and change, time, beliefs, goals, obligations, rationality, etc. In an earlier paper we began this task by abstracting away from many of these features in order to develop a formal theory which captures what we consider to be the essential features of what we call *absolute safety* [9]. In order to make this paper self-contained, we outline the essential features of this theory in the next section. We then extend this theory to include *normative safety*. We begin by introducing *Defeasible Dynamic Logic* and use this to give a possible-worlds semantics for normative safety. We then present a logic of normative safety. We then show how commonsense reasoning about the normal termination of actions can be represented by means of a preferential entailment relation. In conclusion, we discuss the relationships between safety, obligation, rationality and risk, and outline some extensions to the present work.

2 Absolute Safety

We start with disasters and the idea of a *disastrous state*. A disastrous state is one in which some fact, which the agent considers to be disastrous, is true. It is assumed that such states are abhorrent to the agent and that the agent tries to avoid them if at all possible. For instance, a typical medical safety-critical system would consider states in which the patient dies as a result of treatment to be disastrous, and would try to avoid them at all costs. We further assume that the application domain is such that there is no question as to which states the agent considers to be disastrous. Of course, what counts as a disastrous state may be context-dependent. For example, the agent might consider a state in which it has missed a plane to be disastrous if it has a non-transferrable ticket, but it need not do so if there is another flight and the ticket is transferrable. A *disastrous action* can now be defined to be an action which *always* leads to a disastrous state. We then define a *dangerous action* to be an action which *may* lead to a disastrous state, and a *dangerous state* to be state in which every action which the agent can perform – every action which is open to the agent – is dangerous. Note that disastrous actions (states) are thus dangerous actions (states). Finally, we define a *safe action* (a *safe state*) to be an action (a state) which is not dangerous.

We have chosen to formalize safety in Dynamic Logic [6], as this offers a powerful tool for formalizing actions using classical logic and Kripke semantics. In particular, states are represented as possible worlds, and actions are represented as accessibility relations between worlds. The crucial distinction between the necessary and the possible consequences of actions in our informal analysis is thus readily and naturally formalized. We begin by defining the language $\mathcal{SL}$. Let Φ_0 be a set of primitive propositions, and PA be a set of primitive actions. We will use the Roman letters $a, b, \ldots$ (with or without subscripts or superscripts) to denote actions, and lower case Greek letters $\phi, \psi, \ldots$ (with or without subscripts or superscripts) to denote formulas.

Definition 1 (Actions) *The set of (composite) actions* $ACTION$ *is, as usual, defined to be the smallest set which is closed under the following syntactic rules:*

- *If* $a \in PA$ *then* $a \in ACTION$
- *If* $a, b \in ACTION$ *then* $(a \cup b), (a; b) \in ACTION$

Here $\cup$ is the non-deterministic choice operator (so $a \cup b$ means "do either a or b non-deterministically"), and ';' is the sequence operator (so $a; b$ means "do a and then do b").

Definition 2 (The language $\mathcal{SL}$) *The language* $\mathcal{SL}$ *is the minimal set of formulas which satisfies the following conditions:*

- *If* $p \in \Phi_0$ *then* $p \in \mathcal{SL}$.
- *If* $\phi, \psi \in \mathcal{SL}$ *then* $\neg\phi \in \mathcal{SL}$ *and* $\phi \wedge \psi \in \mathcal{SL}$.
- *If* $a \in ACTION$ *and* $\phi \in \mathcal{SL}$ *then* $\langle a \rangle\phi \in \mathcal{SL}$.
- *If* $\phi \in \mathcal{SL}$ *then* $\mathbf{Dis}\phi \in \mathcal{SL}$ *and* $\mathbf{Dan}\phi \in \mathcal{SL}$.
- *If* $a \in ACTION$ *then* $\mathbf{Dis}(a) \in \mathcal{SL}$ *and* $\mathbf{Dan}(a) \in \mathcal{SL}$.

Informally, $\langle a \rangle\phi$ states that doing a makes ϕ possible, that is ϕ *may* be true after action a is taken. Then $[a]\phi$ is defined as $\neg\langle a \rangle\neg\phi$. Thus $[a]\phi$ states that doing a makes ϕ inevitable; that is, ϕ *will* be true after action a is taken. A sentence of the form $\mathbf{Dis}\phi$ ($\mathbf{Dan}\phi$) means that ϕ is disastrous (dangerous). Similarly $\mathbf{Dis}(a)$ ($\mathbf{Dan}(a)$) means that action a is disastrous (dangerous). The safety operators are then introduced by definition:

$$\mathbf{Safe}(a) \stackrel{\text{def}}{\Longleftrightarrow} \neg\mathbf{Dan}(a)$$

$$\mathbf{Safe}\phi \stackrel{\text{def}}{\Longleftrightarrow} \neg\mathbf{Dan}\phi$$

Recall that our analysis of safety started with disastrous states. We therefore extend the standard Kripke models of propositional Dynamic Logic by adding a disastrous-worlds function. In our analysis

[1] Applied Logic Group, Department of Computer Science, Queen Mary and Westfield College, University of London

ECAI 96. *12th European Conference on Artificial Intelligence*
Edited by W. Wahlster
Published in 1996 by John Wiley & Sons, Ltd.

we also want to view actions as being temporally forwards-directed, with each action leading from an earlier world to a set of later worlds. On this view, an action followed by its inverse (e.g. the *pickup* and *putdown* operators in STRIPS) can be conceived as leading to a new world which differs from the original one only in that the time stamp has changed; that is, time has moved on. It also seems natural to require that the underlying model of time is a tree, in which each branch represents a possible outcome of a sequence of actions.

Definition 3 ($\mathcal{SL}$-models) *A* model for $\mathcal{SL}$ *is a tuple* $M = \langle W, \{R^a\}_{a \in PA}, D, V\rangle$ *where*

- W *is a (non-empty) set of possible worlds,*
- $R^a \subseteq W \times W$ *is a binary accessibility relation for each primitive action* $a \in PA$*,*
- $D : W \rightarrow \mathcal{P}(W)$ *assigns to each world* w *the set* $D(w)$ *of worlds which are disastrous with respect to* w*, and*
- $V : \Phi_0 \rightarrow \mathcal{P}(W)$ *is the usual valuation function.*

As usual, we extend the accessibility relations for primitive actions to (composite) actions R^a_+ *(usually written as* R^a *where there is no danger of ambiguity) as follows:*

- $R^a_+ = R^a$ *(where* a *is primitive)*
- $R^{a \cup b}_+ = R^a_+ \cup R^b_+$
- $R^{a;b}_+ = R^a_+ \circ R^b_+$ *(composition of relations)*

Finally, we require that the global accessibility relation $R = \{\langle w, w'\rangle : wR^a_+w'$ *for some* $a\}$ *is a* chronology *– that is, that* R *is a backwards-linear (or anti-convergent) partial order.*

Intuitively wR^a_+w' means that world w' is one possible outcome of doing action a in world w. In order to capture the idea of the consequences of an action *leading to* a disastrous state, we define the *trace* of each action a from each world w. The trace of a from w, is a tree rooted at w, each branch of which represents a course of events which might result from doing a at w. Thus a may lead to a disastrous world (state) if there is such a world on some branch of the trace of a from w. Formally, let $R^a w = \{w' \in W : \langle w, w'\rangle \in R^a_+\}$ be the set of all possible outcomes of doing the (possibly compound) action a at w. Then $Trace(a, w)$ is defined as follows.

- $Trace(a, w) = \langle w, w\rangle \cup \{\langle w, w'\rangle : w' \in R^a w\}$ where a is primitive
- $Trace((a \cup b), w) = Trace(a, w) \cup Trace(b, w)$
- $Trace((a; b), w) = Trace(a, w) \cup \bigcup_{w' \in R^a w} Trace(b, w') \cup (Trace(a, w) \circ \bigcup_{w' \in R^a w} Trace(b, w'))$

For brevity's sake we will write $w_1 R^a w w_2$ if $\langle w_1, w_2\rangle \in Trace(a, w)$; thus, $w_1 R^a w w_2$ states that w_1 occurs no later than w_2 on a branch in the trace of a from w.

Definition 4 (Truth conditions for $\mathcal{SL}$) *Let* $M = \langle W, \{R^a\}_{a \in PA}, D, V\rangle$ *be an* $\mathcal{SL}$*-model. Then a sentence* ϕ *is true at a world* w *in* M *(written* $M, w \models \phi$*, or, equivalently,* $w \in [\![\phi]\!]^M$*) as follows.*

$M, w \models p$	*iff*	$w \in V(p)$ *where* p *is primitive*
$M, w \models \neg\psi$	*iff*	$M, w \not\models \psi$
$M, w \models \psi \wedge \chi$	*iff*	$M, w \models \psi$ *and* $M, w \models \chi$
$M, w \models \langle a\rangle\psi$	*iff*	$\exists w'(w' \in R^a w$ *and* $w' \in [\![\psi]\!]^M)$
$M, w \models \mathbf{Dis}\psi$	*iff*	$[\![\psi]\!]^M \subseteq D(w)$
$M, w \models \mathbf{Dis}(a)$	*iff*	$\forall w_1(wR^a ww_1 \Rightarrow \exists w_2((w_1 R^a ww_2$ *or* $w_2 R^a ww_1)$ *and* $w_2 \in D(w)))$
$M, w \models \mathbf{Dan}(a)$	*iff*	$\exists w'(wR^a ww'$ *and* $w' \in D(w))$
$M, w \models \mathbf{Dan}\psi$	*iff*	$\forall w_1(w_1 \in [\![\psi]\!]^M \Rightarrow (\forall a)(R^a w_1 \neq \emptyset \Rightarrow \exists w_2(w_1 R^a w_1 w_2$ *and* $w_2 \in D(w))))$

As usual, a sentence ϕ *is said to be true in a model* M *(written* $M \models \phi$*) if* $M, w \models \phi$ *for all worlds* w *in* M*, and* ϕ *is said to be valid (written* $\models \phi$*) if* ϕ *is true in all models.*

The logic of absolute safety, **SL**, is given in the companion paper [9], along with a detailed example showing how it can be used to reason about safety in a nuclear power station.

3 Normative Safety

We have defined an *absolute* notion of safety, in which an action is safe only if there is absolutely no possibility of it leading to a dangerous state. In practice, safety is a *normative* concept. That is, in everyday life almost all actions are dangerous, as there is always *some* (perhaps remote) possibility that the action will lead to a disastrous state. So in practice we are interested in actions which are *normally* safe; that is, in actions which do not normally lead to disastrous states. For example, it is not absolutely safe to fly by major airlines (because, there is always the possibility, however remote, of a crash) but it is normally safe to do so (as flights do not normally crash). In this section we extend the theory of safety to reasoning of this kind by adding a normative dimension to the concepts introduced in the previous section. For example, a *normally* dangerous action is one which, in the normal course of events, may lead to a disastrous state, and a *normally* dangerous state is one in which every action which is open to the agent is, in the normal course of events, dangerous.

3.1 The Language $\mathcal{SL_N}$

In order to formalise the normative aspect, we introduce *Defeasible Dynamic Logic* by adding an abnormality operator, Ab, and a defeasible necessity operator, $[\cdot]_N$ to $\mathcal{SL}$.

Definition 5 (The language $\mathcal{SL_N}$) $\mathcal{SL_N}$ *is defined by replacing* $\mathcal{SL}$ *with* $\mathcal{SL_N}$ *everywhere in Definition 2, and by adding the clauses:*

- *If* $a \in ACTION$ *then* $Ab(a) \in \mathcal{SL_N}$*.*
- *If* $a \in ACTION$ *and* $\phi \in \mathcal{SL_N}$ *then* $[a]_N\phi \in \mathcal{SL_N}$*.*
- *If* $\phi \in \mathcal{SL_N}$ *then* $\mathbf{Dis}_{\mathrm{N}}\phi \in \mathcal{SL_N}$ *and* $\mathbf{Dan}_{\mathrm{N}}\phi \in \mathcal{SL_N}$*.*
- *If* $a \in ACTION$ *then* $\mathbf{Dis}_{\mathrm{N}}(a) \in \mathcal{SL_N}$ *and* $\mathbf{Dan}_{\mathrm{N}}(a) \in \mathcal{SL_N}$*.*

Intuitively, $Ab(a)$ means that action a has ended abnormally, $[a]_N\phi$ states that ϕ holds in all normal outcomes of a, $\mathbf{Dis}_{\mathrm{N}}\phi$ ($\mathbf{Dan}_{\mathrm{N}}\phi$) states that ϕ is normally disastrous (normally dangerous), and $\mathbf{Dis}_{\mathrm{N}}(a)$ ($\mathbf{Dan}_{\mathrm{N}}(a)$) states that action a is normally disastrous (normally dangerous). The additional operators are then defined as follows:

$$\langle a\rangle_N\phi \overset{\text{def}}{\iff} \neg[a]_N\neg\phi$$

$$\mathbf{Safe}_{\mathrm{N}}(a) \overset{\text{def}}{\iff} \neg\mathbf{Dan}_{\mathrm{N}}(a)$$

$$\textbf{Safe}_{\text{N}}\phi \overset{\text{def}}{\Longleftrightarrow} \neg\textbf{Dan}_{\text{N}}\phi$$

Thus $\langle a\rangle_N\phi$ states that ϕ holds in some normal outcome of a, and $\textbf{Safe}_{\text{N}}(a)$ ($\textbf{Safe}_{\text{N}}\phi$) states that action a (state ϕ) is normally safe.

$\mathcal{SL}$-models are extended to $\mathcal{SL}_{\mathcal{N}}$-models by adding the function AB which returns the set of actions which have ended abnormally at each world.

Definition 6 ($\mathcal{SL}_{\mathcal{N}}$-models) *An $\mathcal{SL}_{\mathcal{N}}$-model is a tuple $M = \langle W, \{R^a\}_{a\in PA}, AB, D, V\rangle$, where:*

- *W, $\{R^a\}_{a\in PA}$, D and V are as in Definition 3, and*
- *$AB : W \to \mathcal{P}(ACTION)$ assigns to each world w the set of actions which have ended abnormally at w. AB is required to satisfy the following conditions on compound actions:*
  - *$a;b \in AB(w)$ iff $\exists w', w''(w''R^a w'$ and $w'R^b w$ and $b \in AB(w))$, and*
  - *$a \cup b \in AB(w)$ iff $\exists w'(w'R^a w$ and $a \in AB(w)$, or $w'R^b w$ and $b \in AB(w))$.*

The conditions on AB ensure that if an action $a;b$ has ended abnormally at a world w then b has ended abnormally at w, and that if an action $a \cup b$ has ended abnormally at w then one of the actions a or b has ended abnormally at w.

The truth conditions for the normative safety operators are defined in terms of the notion of the *normative trace*, $Trace_N(a, w)$, of an action a from a world w. Informally, $Trace_N(a, w)$ is the subtree of $Trace(a, w)$ consisting of all branches of $Trace(a, w)$ in which events unfold normally, that is, of all branches which contain no worlds in which any sub-action of a ends abnormally. Formally, $Trace_N(a, w)$ is defined analogously to $Trace(a, w)$, except that $R^a_N w = \{w' \in W : \langle w, w'\rangle \in R^a_+$ and $a \notin AB(w')\}$ replaces $R^a w$ everywhere in the definition. In like manner we also abbreviate $\langle w_1, w_2\rangle \in Trace_N(a, w)$ to $w_1 R_N w w_2$.

Definition 7 (Truth conditions for $\mathcal{SL}_{\mathcal{N}}$) *Let $M = \langle W, \{R^a\}_{a\in PA}, AB, D, V\rangle$ be an $\mathcal{SL}_{\mathcal{N}}$-model. Then a sentence ϕ is true at a world w in M (written $M, w \models \phi$, or, equivalently, $w \in [\![\phi]\!]^M$) as in Definition 4 with the following additions.*

$$
\begin{array}{lll}
M, w \models Ab(a) & \textit{iff} & a \in AB(w) \\
M, w \models [a]_N\psi & \textit{iff} & \forall w'(w' \in R^a_N w \Rightarrow w' \in [\![\psi]\!]^M) \\
M, w \models \textbf{Dis}_{\text{N}}\psi & \textit{iff} & M, w \models \textbf{Dis}\psi \\
M, w \models \textbf{Dis}_{\text{N}}(a) & \textit{iff} & \forall w_1(wR^a_N ww_1 \Rightarrow \exists w_2((w_1R^a_N ww_2 \\
 & & \textit{or } w_2R^a_N ww_1) \textit{ and } w_2 \in D(w))) \\
M, w \models \textbf{Dan}_{\text{N}}(a) & \textit{iff} & \exists w'(wR^a_N ww' \textit{ and } w' \in D(w)) \\
M, w \models \textbf{Dan}_{\text{N}}\psi & \textit{iff} & \forall w_1(w_1 \in [\![\psi]\!]^M \Rightarrow (\forall a)(R^a w_1 \\
 & & \neq \emptyset \Rightarrow \exists w_2(w_1R^a_N w_1w_2 \textit{ and} \\
 & & w_2 \in D(w))))
\end{array}
$$

3.2 The Safety Logic $\mathbf{SL_N}$

The logic of (absolute and) normative safety, $\mathbf{SL_N}$, is obtained by adding the following axioms and inference rules to the absolute safety logic **SL** [9].

Axioms

(AN1) $[a]_N\top$
(AN2) $[a]_N(\phi \wedge \psi) \leftrightarrow [a]_N\phi \wedge [a]_N\psi$
(AN3) $[a;b]_N\phi \leftrightarrow [a]_N[b]_N\phi$
(AN4) $[a \cup b]_N\phi \leftrightarrow [a]_N\phi \wedge [b]_N\phi$
(AN5) $[a]_N\phi \leftrightarrow [a](\neg Ab(a) \to \phi)$

(AB1) $Ab(a;b) \to Ab(b)$
(AB2) $Ab(a \cup b) \leftrightarrow Ab(a) \vee Ab(b)$

(DN1) $\textbf{Dis}_{\text{N}}\phi \leftrightarrow \textbf{Dis}\phi$
(DN2) $\textbf{Dis}_{\text{N}}(a \cup b) \leftrightarrow \textbf{Dis}_{\text{N}}(a) \wedge \textbf{Dis}_{\text{N}}(b)$

(DAN1) $\textbf{Dan}_{\text{N}}\phi \wedge \textbf{Dan}_{\text{N}}\psi \to \textbf{Dan}_{\text{N}}(\phi \wedge \psi)$
(DAN2) $\langle a\rangle_N\phi \wedge \textbf{Dis}_{\text{N}}\phi \to \textbf{Dan}_{\text{N}}(a)$
(DAN3) $\textbf{Dan}_{\text{N}}(a \cup b) \leftrightarrow \textbf{Dan}_{\text{N}}(a) \vee \textbf{Dan}_{\text{N}}(b)$
(DAN4) $\textbf{Dan}_{\text{N}}(a) \vee (\langle a;b\rangle_N\phi \wedge \textbf{Dan}_{\text{N}}\phi) \to \textbf{Dan}_{\text{N}}(a;b)$

(DDN1) $\textbf{Dis}_{\text{N}}\phi \to \textbf{Dan}_{\text{N}}\phi$
(DDN2) $\textbf{Dis}_{\text{N}}(a) \to \textbf{Dan}_{\text{N}}(a)$

(DDAN1) $\textbf{Dis}(a) \to \textbf{Dis}_{\text{N}}(a)$
(DDAN2) $\textbf{Dan}_{\text{N}}(a) \to \textbf{Dan}(a)$
(DDAN3) $\textbf{Dan}_{\text{N}}\phi \to \textbf{Dan}\phi$

Definitions

(PosN) $\langle a\rangle_N\phi \leftrightarrow \neg[a]_N\neg\phi$
(SAdf) $\textbf{Safe}_{\text{N}}(a) \leftrightarrow \neg\textbf{Dan}_{\text{N}}(a)$
(SSdf) $\textbf{Safe}_{\text{N}}\phi \leftrightarrow \neg\textbf{Dan}_{\text{N}}\phi$

Inference Rules

(MONAN) From $[a]_N\phi$ and $\phi \to \psi$ infer $[a]_N\psi$
(SPDSN) From $\psi \to \phi$ infer $\textbf{Dan}_{\text{N}}\phi \to \textbf{Dan}_{\text{N}}\psi$

Theorems and Derived Rules

(PAN) $\langle a\rangle_N\phi \leftrightarrow \langle a\rangle(\phi \wedge \neg Ab(a))$
(SPSN) From $\phi \to \psi$ infer $\textbf{Safe}_{\text{N}}\phi \to \textbf{Safe}_{\text{N}}\psi$
(SSN1) $\neg\textbf{Safe}_{\text{N}}\bot$
(SSN2) $\textbf{Safe}_{\text{N}}(\phi \wedge \psi) \leftrightarrow \textbf{Safe}_{\text{N}}\phi \wedge \textbf{Safe}_{\text{N}}\psi$
(SSN3) $\textbf{Safe}_{\text{N}}(\phi \wedge \psi) \to \textbf{Safe}_{\text{N}}\phi \vee \textbf{Safe}_{\text{N}}\psi$
(SAN1) $\textbf{Safe}_{\text{N}}(a \cup b) \leftrightarrow \textbf{Safe}_{\text{N}}(a) \wedge \textbf{Safe}_{\text{N}}(b)$
(SAN2) $\textbf{Safe}_{\text{N}}(a;b) \to \textbf{Safe}_{\text{N}}(a)$

The axioms (AN1)-(AN5), (AB1), (AB2), (PosN) and the inference rule (MONAN) give our Defeasible Dynamic Logic. The remaining axioms and inference rule state properties of the defeasible safety operators and the relationship between them and the absolute safety operators. Some of the properties of normally safe actions and sates are listed as theorems.

Theorem 1 (Soundness of $\mathbf{SL_N}$) *The logic $\mathbf{SL_N}$ is sound for the class of $\mathcal{SL}_{\mathcal{N}}$-models.*

It is not possible to prove the completeness of $\mathbf{SL_N}$ by means of the standard method of canonical models. In on-going work we are looking at other methods and at proving the independence of the axioms.

Example 1 *Let Θ be the theory:*

$$\textbf{Safe}(fly) \leftrightarrow [fly]Arrive \tag{1}$$
$$\textbf{Safe}_{\text{N}}(fly) \leftrightarrow [fly]_N Arrive \tag{2}$$
$$Preconds(fly) \to$$

$$[fly]_N Arrive \wedge \langle fly\rangle Arrive \wedge \langle fly\rangle \neg Arrive \quad (3)$$

$$Preconds(fly) \quad (4)$$

$$Ab(fly) \leftrightarrow Crashed \vee Hijacked \vee \ldots \quad (5)$$

That is, the flight is absolutely safe if, no matter what happens, it arrives at the destination (1), the flight is normally safe if it arrives at the destination in the normal course of events (2), if the usual preconditions hold then, in the normal course of events, the flight will arrive at the destination but there is always the possibility that it does not (3), the usual preconditions hold (4), and, finally the flight proceeds normally if the plane does not crash, is not hijacked, etc. (5). Then $\Theta \models \mathbf{Safe}_N(fly) \wedge \neg\mathbf{Safe}(fly)$. *That is, given* Θ *we can infer that it is normally safe to fly, but not absolutely safe to do so. However, we cannot infer that it is possible for the flight to arrive safely. As is clear from (AN5), the conclusion* $[fly]_N Arrive$ *only guarantees arrival whenever the flight ends normally, it does not guarantee that the flight does ever end normally.*

3.3 Chronologically Minimising Abnormalities

While it is now possible to distinguish between absolute safety and normative safety and to reason about normative safety, we still do not have a complete account of commonsense reasoning about it. For example, in reasoning that it is normally safe to fly, we make the default assumption that, given that the usual preconditions hold (the plane is airworthy, has been fuelled, etc.), the flight will be a normal one. It should thus at least be possible that the flight ends normally. In order to represent this aspect of commonsense reasoning about actions, we will, following [11], define a preferential entailment relation. Intuitively, we prefer models in which events unfold normally. Technically, we restrict consideration to models in which abnormalities occur as late as possible; that is, to models which are chronologically most normal.

Definition 8 *Let* $M = \langle W, \{R^a\}_{a\in PA}, AB, D, V\rangle$ *and* $M' = \langle W', \{R^{a'}\}_{a\in PA}, AB', D', V'\rangle$ *be* $\mathcal{SL}_{\mathcal{N}}$*-models with respective chronologies* R *and* R'. M *and* M' *are said to be* comparable *if* $W = W'$, $R = R'$ *and* $D = D'$. M *is said to be* chronologically more normal than M' *(written* $M \prec M'$*) if* M *and* M' *are comparable and there is a world* $w \in W$ *such that:*

- *for any world* $w' \in W$ *such that* $w \neq w'$ *and not* wRw',
  - $AB(w') = AB'(w')$ *and*
  - $\{p \in \Phi_0 : w' \in V(p)\} = \{p \in \Phi_0 : w' \in V'(p)\}$,
- *for any action a,* $M', w \models Ab(a)$ *if* $M, w \models Ab(a)$*, and*
- *for some action a,* $M', w \models Ab(a)$ *and* $M, w \not\models Ab(a)$.

Thus $M \prec M'$ if M and M' are comparable, M and M' agree on all worlds which precede or are unrelated to some world w in their common chronology, every action that ends abnormally at w in M also ends abnormally at w in M', and there is at least one action which ends normally at w in M and which ends abnormally at w in M'.

Definition 9 *An* $\mathcal{SL}_{\mathcal{N}}$*-model* M *is a* chronologically most normal model *(a* c.m.n.-model*) of a sentence* ϕ *if* $M \models \phi$ *and there is no model* M' *such that* $M' \models \phi$ *and* $M' \prec M$. *Similarly,* M *is a c.m.n.-model of a set of sentences* Θ *if* $M \models \Theta$ *and there is no model* M' *such that* $M' \models \Theta$ *and* $M' \prec M$.

Definition 10 (Normal entailment) *A set of sentences* Θ normally entails *a sentence* ϕ *(written* $\Theta \approx\!\!\!\!\models \phi$*) if* $M \models \phi$ *for any c.m.n.-model* M *of* Θ.

Example 2 *Let* Θ *be as in Example 1. Then:* $\Theta \approx\!\!\!\!\models \mathbf{Safe}_N(fly) \wedge \neg\mathbf{Safe}(fly) \wedge \langle fly\rangle_N Arrive$. *Given* Θ*, we can conclude that it is normally safe to fly, but not absolutely safe to do so, and, moreover, we can also conclude by default that it is possible that the flight ends normally; that is, that there is at least one outcome of the flying action which does not involve an abnormality. (In outline, the proof is as follows. By chronological minimisation of abnormalities we have* $\neg[fly]Ab(fly)$*, that is* $\langle fly\rangle\neg Ab(fly)$*. This together with* $[fly]_N Arrive$*, that is,* $[fly](\neg Ab(fly) \rightarrow Arrive)$*, (AN5), gives* $\langle fly\rangle(\neg Ab(fly) \wedge Arrive)$*, that is* $\langle fly\rangle_N Arrive$*, (PAN).)*

A general argument for the use of chronological minimisation in *predictive* commonsense reasoning is given in [1, 2].

4 Safety, Obligation, Rationality and Risk

In this section we discuss some the relationships between safety, obligation, rationality and risk, and indicate how the present work might be extended to include these.

It is interesting to consider the relationship between obligation and safety. A central idea here is that of a *safety procedure* or *protocol*. Protocols are designed to maintain safety and have to be revised if they are found to lead to dangerous states. Protocols impose deontic restrictions on agents, as the agents involved are (normally) obliged to follow them. For example, in hospitals medical staff are (normally) obliged to follow protocols when treating patients; see, e.g, [5]. In order to formalise these constraints we could add the deontic operators $\mathbf{Obl}$ and $\mathbf{Per}$ to $\mathcal{SL}_{\mathcal{N}}$. Intuitively, $\mathbf{Obl}(a)$ states that action a is obligatory, and $\mathbf{Per}(a)$ states that action a is permitted. Of course we have to provide appropriate formal semantics for these operators. One trick might be to introduce an action operator, Do, and then use the standard possible-world semantics for deontic operators; for example, $\mathbf{Obl}(a) \stackrel{\text{def}}{\Longleftrightarrow} \mathbf{Obl}Do(a)$. We can then, for example, state that only procedures which are normally safe can be followed:

$$\mathbf{Per}(a) \rightarrow \mathbf{Safe}_N(a).$$

The problem with this approach is that it may not always be *rational* to do only safe actions. For example if a patient has reached a critical stage, it may be that an *unsafe* treatment is the only option. So, in exceptional circumstances agents must be able to override protocols. Moreover, in domains where safety is not considered critical, it is often rational to do things which are not even normally safe. For example, an investor may risk money on an unsafe investment as this promises to bring in a greater return than any safe alternative, or an adult may risk their life trying to save that of a child. Indeed, in much of everyday life it is rational to take risks; that is to do dangerous things. In the case of protocols, we thus want the imperative to sate that only rational actions are permitted:

$$\mathbf{Per}(a) \rightarrow \mathbf{Rat}(a).$$

The problem now is to give a semantic account of the rationality operator. One starting point is the AI-planning theory of practical rationality outlined in [3] which begins with the following definition:

> A resource-bounded agent behaves rationally if it reasons and acts so as to achieve as many of its goals, in their comparative order of importance to the agent, as is possible given the resources available to it and the constraints in force.

In this paper we will simplify the theory dramatically. The relevant components are an agenda of goals, a planner which produces plans for these goals, and a scheduler which chooses the best plan to execute.

The agenda can be formalised using the preference semantics developed in [7, 8]. In order to incorporate them, two components need to be added to our models:

- $cw : W \times \mathcal{P}(W) \to \mathcal{P}(W)$ is a function, which selects the set $cw(w, [\![\phi]\!]^M)$ of closest worlds to w in which ϕ is true, and
- $\succ \subseteq \mathcal{P}(W) \times \mathcal{P}(W)$ is a preference relation on sets of worlds.

We can then introduce the preference operator $\mathbf{P}$ and give its truth conditions as follows:

$$M, w \models \phi\mathbf{P}\psi \text{ iff } cw(w, [\![\phi \wedge \neg\psi]\!]^M) \succ cw(w, [\![\psi \wedge \neg\phi]\!]^M).$$

So $\phi\mathbf{P}\psi$ is true at w if every closest $\phi \wedge \neg\psi$-world to w is preferred to every closest $\psi \wedge \neg\phi$-world to w. The truth condition thus formalises von Wright's *conjunction expansion principle* for preferences. We can now formalise the agenda by writing $\phi\mathbf{P}\phi'$ to indicate that goal ϕ is preferred to goal ϕ'.

In order to represent the planner we will, for present purposes, simplify and assume that there is only one possible plan (action) which will achieve each goal. We can therefore define the plan a which will achieve goal ϕ as:

$$Plan(a, \phi) \overset{\text{def}}{\Longleftrightarrow} [a]\phi \wedge \forall b(b \neq a \to \neg[b]\phi).$$

We can then define a preference ordering on plans according to the desirability of the goals they are designed to achieve:

$$a \succ_d b \overset{\text{def}}{\Longleftrightarrow} Plan(a, \phi) \wedge Plan(b, \psi) \wedge \phi\mathbf{P}\psi$$

A plan (action) can then be defined to be executable if it is possible given the resources available: $Exec(a) \overset{\text{def}}{\Longleftrightarrow} \neg[a]\bot$. The scheduler should then execute the most preferred executable plan:

$$\mathbf{Rat}(a) \overset{\text{def}}{\Longleftrightarrow} Exec(a) \wedge \neg\exists b(Exec(b) \wedge b \succ_d a).$$

It may be important to reason about degrees of risk; for example, an investor should take reasonable risks. In order to represent this kind of reasoning, the operator Sat can be added to the language, where, intuitively, $Sat(a)$ means that the risk associated with action a is acceptable. Models can be extended to include a function which associates an appropriate level of risk with each action at each world:

$$SR : W \times ACTION \to \omega$$

Then, letting $Risk(a, w)$ be the number of branches in the (normative) trace of a from w which contain at least one disastrous world, we can supply truth conditions for the new operator as follows:

$$M, w \models Sat(a) \quad \textit{iff} \quad Risk(a, w) \leq SR(a, w)$$

The most preferred, executable action could then be defined to be rational if the degree of risk involved is acceptable:

$$\mathbf{Rat}(a) \overset{\text{def}}{\Longleftrightarrow} Exec(a) \wedge \neg\exists b(Exec(b) \wedge b \succ_d a) \wedge Sat(a).$$

5 Concluding Remarks

In this paper we extended our analysis and formalisation of absolute safety to normative safety. We extended the formalism by introducing a defeasible extension of dynamic logic in order to give possible-worlds semantics and a logic for normative safety. We then defined a preferential entailment relation in order to be able to represent commonsense reasoning about actions. We then discussed the relationship between safety, obligation and rationality, and gave an indication of how the present work might be extended to produce a formal theory of safety, obligation, rationality, and risk.

As a result of the review process we have become aware of the work on reasoning about safety in theoretical Computer Science, for example that of [10], and of Dunin-Keplicz and Radzikowska's extension of Dynamic Logic to incorporate actions with typical effects [4].

ACKNOWLEDGEMENTS

This research formed part of the RED project, *Decision Support Systems, Safety and Liability*, and was supported by the United Kingdom Engineering and Physical Sciences Research Council under grant number GR/H 12348. We would also like to thank the reviewers and everyone else who has commented on it.

REFERENCES

[1] Bell, J., Prediction Theories and Explanation Theories. Unpublished manuscript. Available from ftp.dcs.qmw.ac.uk in a directory named applied_logic/bell as a file pt&et.ps.

[2] Bell, J., Remarks on the Evaluation of Formal Theories of Causal Reasoning. Working Notes of the IJCAI-95 Workshop on *Nonmonotonic Reasoning, Action and Change*, Williams, M., (ed.), pp. 14-21.

[3] Bell, J., A Planning Theory of Practical Rationality. Proceedings of the AAAI-95 Fall Symposium on Rational Agency, M.I.T., 1995. Fehling, M. (ed.), pp. 1-4.

[4] Dunin-Keplicz, B., and Radzikowska, A., Epistemic Approach to Actions with Typical Effects. Proceedings of *ECSQARU'95*. Froidevaux, C., and Kholas, J., (Eds). Lecture Notes in Artificial Intelligence No. 946. Springer, Berlin, 1995, pp. 180-188.

[5] Hammond, P., and Sergot, M., Computer Support for Protocol-Based Treatment of Cancer. Proceedings of the 2nd International Conference on the Practical Applications of Prolog, London, 1994.

[6] Harel, D., Dynamic Logic, in: D. Gabbay and F. Guenthner, (eds.), *Handbook of Philosophical Logic*, Vol.II, (D. Reidel publishing company, 1984), 497-604.

[7] Huang, Z., *Logics for Agents with Bounded Rationality*, ILLC Dissertation series 1994-10, University of Amsterdam, 1994.

[8] Huang, Z., Masuch, M., and Pólos, L., ALX: an action logic for agents with bounded rationality, *Artificial Intelligence* **82** (1996), pp. 101-153.

[9] Huang, Z., and Bell, J., Safety Logics I : Absolute Safety. Proceedings of Common Sense '96. Buvac, S., and Costello, T., (eds.), pp. 59-66.

[10] Manna, Z., and Pneuli, A., *The Temporal Logic of Reactive and Concurrent Systems: Specification*, Springer Verlag, New York, 1991.

[11] Shoham, Y. *Reasoning About Change*, M.I.T. Press, Cambridge, Massachusetts, 1988.

[12] von Wright, G., *The Logic of Preference*, (Edinburgh, 1963).