Department of Computer Science,
University of Otago

UNIVERSITY of OTAGO

SAPERE AUDE

Te Whare Wānanga o Otāgo

Technical Report OUCS-2001-09

A Preferential Semantics for Epistemic Logic

Authors:
David Ferguson
Willem Labuschagne

Status: submitted to AAMAS 2002

Department of Computer Science,
University of Otago, PO Box 56, Dunedin, Otago, New Zealand

http://www.cs.otago.ac.nz/trseries/

STUDENT PAPER: ID 122: A Preferential Semantics for Epistemic Logic

Dave Ferguson
University of Otago
Dunedin, New Zealand
dferguso@cs.otago.ac.nz

Willem Labuschagne
University of Otago
Dunedin, New Zealand
willem@cs.otago.ac.nz

ABSTRACT

The development of agent communication languages casts a spotlight on epistemic logic and the enrichment of epistemic languages by additional operators, e.g. deontic operators or operators representing speech acts. In this paper we focus on two limitations of classical epistemic logic. The standard possible worlds semantics allows one to model *either* the knowledge *or* the beliefs of an agent, but it is not so easy to model both in a manner compatible with the intuition that knowledge shares the same 'dimension' as belief. Furthermore, the distinction between knowledge and belief is intimately tied up with defeasibility and non-monotonicity, which in turn (via total preorders) is connected with epistemic entrenchment. We therefore introduce a generalisation of the possible worlds semantics which not only accommodates knowledge and belief simultaneously but admits a hierarchy of belief operators reflecting different levels of entrenchment.

General Terms

Epistemic logic, knowledge, belief

Keywords

Possible worlds, epistemic entrenchment, defeasibility

1. INTRODUCTION

The explosion of activity following the publications [26, 6] has demonstrated the usefulness of epistemic logic as a core language for the development of more specialised agent communication languages. Current emphasis is on the incorporation of suitable explications of concepts, such as 'trust'. It would seem, however, to be an error to focus so narrowly on ways to enrich the core language that attempts to improve the core language cease. In particular, there is as yet no standard way to represent both an agent's knowledge and the agent's less emphatic beliefs.

AAMAS 2002 Bologna, Italy

Under the standard possible worlds semantics, an interpretation is a triple (S, R, V) in which S is a set of possible worlds, $R \subseteq S \times S$ is an accessibility relation, and V is some function determining which atomic propositions are satisfied relative to which possible worlds.

This semantics supports a corresponding modal operator K in the language: for every proposition ϕ, $K\phi$ is satisfied relative to $s \in S$ if and only if ϕ is satisfied relative to s' for all $s' \in S$ such that $(s, s') \in R$. Depending on the restrictions imposed on R, K may represent either 'knowledge' or 'belief'. In order to support both a knowledge operator K and a belief operator, say B, it would be necessary to add a second accessibility relation. The two operators would thus be independent, contrary to the intuition that knowledge lies on the same spectrum as more tentative beliefs.

If we are to accommodate both knowledge and belief, then it is natural to think of the beliefs as defeasible, and to seek to support such beliefs by semantic structures such as the orderings on S discussed in [15, 16, 21]. In the following sections we describe a way to model an agent whose beliefs may differ in the tentativeness or conviction with which they are believed. Knowledge will arise as a special case. The distinction between knowledge and more tentative forms of belief reduces to a difference between definite and indefinite information.

2. INFORMATION

Let S be the set of all valuations of some propositional language. The notion of information associated with Shannon's work in communication theory suffers from the failure to take into account the meaning of messages, as pointed out in [1]. The theory of semantic information corrects this fault by taking the semantic content of a proposition ϕ to be determined by the worlds "excluded" by ϕ, i.e. relative to which ϕ is not satisfied [2, 12]. Indeed, the division of the set S of worlds into complementary subsets C of worlds that are "included" and C' of worlds that are "excluded" is logically fundamental, for if S is infinite there are subsets C such that for no proposition ϕ is C the set of all worlds at which ϕ is satisfied. An agent may have definite information, in the form of a division of S into C and C', which the agent may or may not be capable of expressing as a proposition. For simplicity we shall assume S to be finite, so that for every subset C' of "excluded" worlds there is some proposition ϕ such that ϕ excludes C', i.e. such that ϕ is satisfied at s if and only if $s \notin C'$.

Intuitively, agents may derive information from sensors (or even from other agents) which is definite. One thinks

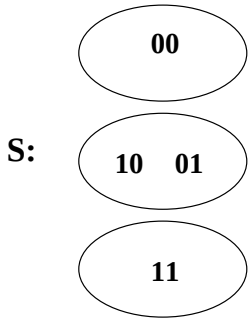

Figure 1: A representation of a default rule in the light-fan system as an ordering on worlds. The status of each component is either on (specified by a 1 in the component's position) or off (specified by a 0 in the component's position). Thus, world 10 is the state of the system where the light is on and the fan is off.

of the impact of a toe against a rock as imparting definite information about the existence of an obstacle. However, agents may also acquire indefinite information (default rules). Whereas definite information may be represented semantically by a division of S into a complementary pair of subsets C and C', default rules may conveniently be represented by order relations on S [7, 25, 15, 16]. The order relations on S that represent default rules may be taken to be either total preorders (as in the belief revision literature, e.g. [13, 9, 3]) or strict modular partial orders (e.g. [16, 22]). The essential feature of such orderings is that the set S is divided into layers. Intuitively, the bottom layer consists of the worlds that are most typical or preferred, and higher layers are increasingly less typical or less preferred.

For example, assume an agent inhabits the light-fan system. This system has two components: a light and a fan. Both components may be on or off. Now, a possible default rule may be that both the light and the fan are usually on, and very seldom are both components off. Such a default rule could be expressed as an ordering of the four possible worlds of the system into 3 layers. The bottom layer would hold the world where both components are on. The middle layer would hold the two worlds where only one component is on. The top layer would hold the world where neither component is on (see Figure 1).

Given the semantic representations of definite and of indefinite information, how are these to be reconciled? One possibility is to use a templated ordering of S. By a templated ordering of S we understand the following. Suppose S has n members. A template for S is any chain with exactly $n+1$ members, say $T = \{0, 1, \ldots, n\}$ under the usual ordering $\leq$. A templated ordering of S is an association of levels in the template with members of S, for example as accomplished by a function $f : S \to T$. All worlds in S are assigned to some level, but there will be levels having no worlds. Definite information is represented by a templated ordering assigning the excluded worlds (those in C') to level n, and the remainder to level 0. Indefinite information is represented by a templated ordering assigning worlds to levels $0, 1, \ldots, n-1$ only.

The reconciliation of definite and indefinite information is achieved by lexicographic refinement in the following sense.

Definition 1. Let $f_d : S \to T$ be the templated ordering representing an agent's definite information and let $f_i : S \to T$ represent the indefinite information. Let f_{d+i} be given by

$$f_{d+i}(s) = \begin{cases} f_d(s) & \text{if } s \in C' \\ f_i(s) & \text{otherwise.} \end{cases}$$

Thus f_{d+i} places the excluded worlds in level n, and pushes tentatively towards exclusion those worlds that, according to the default rule, are less preferred. One may think of f_{d+i} as 'partitioning' S into $n+1$ sets of worlds, some of which are empty, and linearly ordering these $n+1$ sets.

3. GENERALISING POSSIBLE WORLDS SEMANTICS

The method of representing definite knowledge by partitioning the set S into two sets, C and C', is directly equivalent to the standard possible worlds framework that uses a binary accessibility relation. Given an accessibility relation, and given $s \in S$, let $C = \{s' : (s, s') \in R\}$, where C is the set of possible candidates for being the actual world s and the complement C' is the set of excluded worlds. Given, relative to each $s \in S$, a partition of S into two sets C and C', an accessibility relation R may be regained simply by reversing the above process.

Now consider extending the process via templated orderings. The effect is to separate S into a whole range of levels, allowing representation of default rules such as that displayed in Figure 1.

Suppose given, at each $s \in S$, a templated ordering f of S into $n+1$ sets of states, some of which may be empty. Intuitively the top level $f^{-1}(n)$ represents the set C' of excluded worlds while the union of the other levels is C, arranged into sets of worlds of different typicality or likelihood. Let us label these sets from bottom to top as $\{P_0, P_1, \ldots, P_n\}$, where $P_0 = f^{-1}(0)$ denotes the set of most likely worlds and $P_n = f^{-1}(n)$ denotes C' (see Figure 2).

Corresponding to each level P_t (where $0 \leq t \leq n$), we may introduce a defeasible belief operator ∇_t whose semantic underpinning is expressed by the requirement that, for any proposition ϕ, $\nabla_t \phi$ is satisfied relative to world s if and only if ϕ is satisfied at all worlds s' such that, in the templated ordering at s, $s' \in P_u$ for some u with $u \leq t$. So $\nabla_t \phi$ holds at s if ϕ holds at all worlds of all levels at or below P_t in the ordering associated with s. Intuitively, ∇_{n-1} represents the knowledge operator, while ∇_0 represents the most tentatively held form of belief, as made precise in the discussion of epistemic entrenchment in Section 5 below.

We typically denote the level P_i in the ordering at world s as P_i^s. This extension from a dichotomous partition of S to a templated ordering of S slightly alters the notion of interpretation. We call the resulting generalisation the system *Not Only Knowing*, or NOK.

Definition 2. A **NOK interpretation** of a modal language with defeasible belief operators $\nabla_0, \ldots, \nabla_n$ is a triple (S, F, V) such that $S = \{s_i : i \leq n\}$ is the set of all valuations of the underlying non-modal language, V is any function expressing the natural relationship between atomic propositions and the valuations that satisfy them, and $F : S \to T^S$ is a function associating with each $s \in S$ a templated ordering of S.

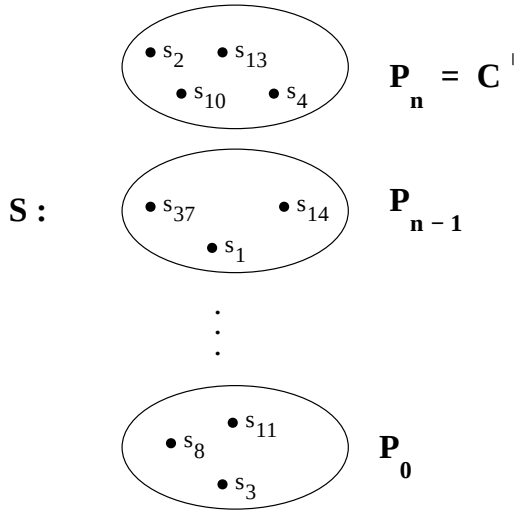

Figure 2: A templated ordering of a set of worlds $S = \{s_1, s_2, \ldots, s_n\}$. All worlds in S are in one of the sets P_0 through P_n. Under the new framework, an agent would construct an ordering like this at each state $s_i \in S$.

Definition 3. Given a NOK interpretation $\mathcal{I} = (S, F, V)$ and a world $s \in S$, $(\mathcal{I}, s) \Vdash \psi$ iff one of the following is the case:

- ψ is an atomic proposition and the valuation $V(s)$ satisfies ψ
- ψ is a boolean combination of shorter propositions and is satisfied according to the usual truth functional criteria
- $\psi = \nabla_t \phi$ and $(\mathcal{I}, s')$ satisfies ϕ for all $s' \in S$ such that $(F(s))(s') \leq t$.

The definitions may readily be generalised; the set of valuations need not be finite, and the set S of possible worlds need not be the set of valuations (in which case V may be taken to be a function from S to the set of valuations, thereby determining which atomic propositions hold at which worlds). If S is infinite, the template for S is $\{0, 1, \ldots, \alpha\}$ where α is the least successor ordinal greater than $card(S)$.

We saw that an accessibility relation can induce and be recovered from an ordering function F that assigns to each world $s \in S$ a dichotomous partition of S into a set C of accessible and a set C' of excluded worlds. More generally, from any ordering function F which assigns to every $s \in S$ a templated ordering of S, an accessibility relation may be derived. For each state s, the top level in the ordering $F(s)$ represents all the states which are inaccessible from s, so we label this set C', then take the union of all levels below this and label the result C.

While this is one way to retrieve the standard epistemic framework from the richer semantics, an equivalent alternative is to ignore all defeasible operators other than ∇_{n-1}, which is taken as the K operator.

The defeasible belief operators have properties compatible with our intuitions regarding degrees of belief. In formulating these we will write $\phi \models \psi$ to assert that in some understood interpretation ψ is satisfied relative to every world s at which ϕ is satisfied, and we write $\Vdash \phi$ to indicate that ϕ is satisfied at every world in every NOK interpretation.

The first property is a consequence of the hierarchical structure of the orderings and holds for every NOK interpretation:

Property 1. $\forall\, t, u : 0 \leq u \leq t \leq n,\ \nabla_t \phi \models \nabla_u \phi$.

In particular, if an agent knows a sentence ϕ, then the agent believes ϕ also.

Another property which holds for every NOK interpretation is the following:

Property 2. $\forall\, t : 0 \leq t \leq n,\ \Vdash \neg \nabla_t false$.

This ensures that an agent will never believe, even tentatively, that which is contradictory.

4. COMPARISON WITH S5

The generalised semantics allows the formulation of counterparts to the classical **S5** properties in terms of which distinctions can be made that are conflated in the standard possible worlds framework.

The **S5** properties are:

- **K** : $(K\phi \wedge K(\phi \rightarrow \alpha)) \rightarrow K\alpha$
- **T** : $K\phi \rightarrow \phi$
- **4** : $K\phi \rightarrow KK\phi$
- **5** : $\neg K\phi \rightarrow K\neg K\phi$.

The Distribution Axiom **K** has a natural translation:

- $\mathbf{K}'$: $(\nabla_i \phi \wedge \nabla_i(\phi \rightarrow \alpha)) \rightarrow \nabla_i \alpha$.

Thus, if an agent defeasibly believes ϕ at s, and defeasibly believes $\phi \rightarrow \alpha$ to the same degree at s, then that agent will defeasibly believe α at s also (again, to the same degree). This property holds for any NOK ordering function F, and thus for any NOK interpretation.

The Truth of Knowledge Axiom **T** has more than one interesting counterpart. The first involves simply rewriting the knowledge operator, K, as its equivalent NOK defeasible operator, ∇_{n-1}, to give:

- $\mathbf{T}'$: $\nabla_{n-1} \phi \rightarrow \phi$.

This property expresses the same idea as the classical **T** property — if an agent knows ϕ at world s then ϕ is true at world s. To impose this property on a frame $\mathcal{F}$ of NOK interpretations we must ensure that, given any world s and ordering P^s at s, we have $s \notin P^s_n$.

A much stronger constraint imposes the Truth of Knowledge property over *all* the defeasible belief operators:

- $\mathbf{T}''$: $\nabla_i \phi \rightarrow \phi$, for every $i < n$.

For this to hold, not only must the agent consider the actual world accessible, but the actual world must always be in the lowest level of the ordering. Thus at each world $s \in S$ with respective ordering P^s, we must have $s \in P^s_0$. Whereas $\mathbf{T}'$ characterises the class of agents with infallible sensors, $\mathbf{T}''$ characterises the class of agents with infallible default rules.

The Positive Introspection Axiom **4** also has more than one interesting counterpart.

A faithful translation replaces all K's by ∇_{n-1}:

- $\mathbf{4}'$: $\nabla_{n-1}\phi \rightarrow \nabla_{n-1}\nabla_{n-1}\phi$.

This states that an agent knows that it knows, yet makes no claims as to whether an agent knows that it defeasibly believes. An alternative is to have each agent knowing that it defeasibly believes, as well as that it knows:

- $\mathbf{4}''$: $\nabla_i\phi \rightarrow \nabla_{n-1}\nabla_i\phi$, for every $i < n$.

Note that $\mathbf{4}''$ implies, in particular:

$\nabla_0\phi \rightarrow \nabla_{n-1}\nabla_0\phi$.

A third alternative is to impose weaker forms of defeasible introspection, for instance requiring that if an agent defeasibly believes a sentence ϕ to some degree, then it defeasibly believes (to the same degree) that it defeasibly believes ϕ:

- $\mathbf{4}'''$: $\nabla_i\phi \rightarrow \nabla_i\nabla_i\phi$.

The weakest form of introspection would seem to be characterised by the following:

- $\mathbf{4}''''$: $\nabla_{n-1}\phi \rightarrow \nabla_0\nabla_{n-1}\phi$.

To convey the flavour of the effect of imposing such constraints, we examine $\mathbf{4}''$ further. What sort of frame would possess property $\mathbf{4}''$?

THEOREM 1. *Let $\mathcal{F}$ be a frame of NOK interpretations (S, F, V). Then $\mathcal{F} \Vdash \nabla_i\phi \rightarrow \nabla_{n-1}\nabla_i\phi$ iff for all $s, t \in S$ it is the case that if $t \notin P_n^s$ then either*

1. *$s \notin P_n^t$ and $P^t = P^s$,* or
2. *$s \in P_n^t$ and for all $u \in S$, if $u \in P_x^s$ and $u \in P_y^t$ then $x \leq y$.*

Thus, if s and t are accessible from each other, they have the same templated orderings. This is a severe constraint, but not unexpectedly so if we think of schema $\nabla_0\phi \rightarrow \nabla_{n-1}\nabla_0\phi$ as expressing such powers of introspection that the agent, for even its most tentatively entertained belief ϕ, *knows* absolutely that it believes ϕ. We would expect a schema at the opposite end of the spectrum, such as $\mathbf{4}''''$ $(\nabla_{n-1}\phi \rightarrow \nabla_0\nabla_{n-1}\phi)$, to impose a much less severe constraint on agents. And so it turns out. By an argument parallel to the above, it is possible to show that $\mathbf{4}''''$ requires only that s and t must have the same sets P_0 and P_n if $s \in P_0^t$ and $t \in P_0^s$.

The second case dealt with in the theorem concerns the possibility that world t is accessible from s but not vice-versa, under which circumstances the ordering at world t must not place any worlds lower than their corresponding positions in the ordering at world s. This is to ensure that anything believed at s is believed to at least the same degree at t.

Analogs of the Negative Introspection Axiom **5** follow a similar pattern, delivering amongst others:

- $\mathbf{5}''$: $\neg\nabla_i\phi \rightarrow \nabla_{n-1}\neg\nabla_i\phi$, for every $i < n$.

THEOREM 2. *Let $\mathcal{F}$ be a frame of NOK interpretations (S, F, V). Then $\mathcal{F} \Vdash \neg\nabla_i\phi \rightarrow \nabla_{n-1}\neg\nabla_i\phi$ iff for all $s, t \in S$ it is the case that if $t \notin P_n^s$ then either*

1. *$s \notin P_n^t$ and $P^t = P^s$,* or

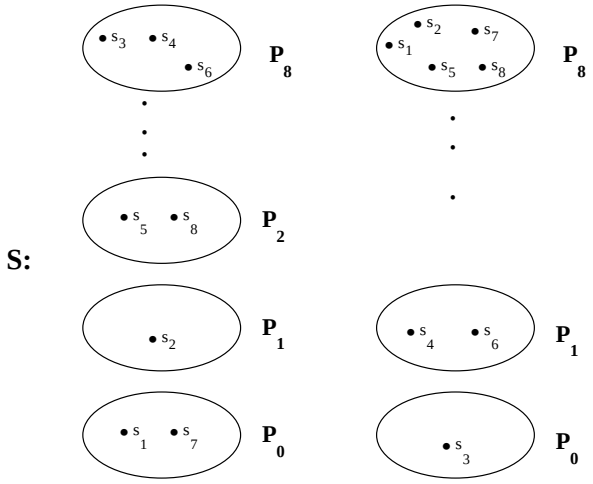

Figure 3: An example NOK ordering function over the set of worlds $S = \{s_1, s_2, \ldots, s_8\}$ which exhibits the $\mathbf{K}'\mathbf{T}'\mathbf{4}''\mathbf{5}''$ properties. The worlds $\{s_1, s_2, s_5, s_7, s_8\}$ all share the same NOK ordering, as do the states $\{s_3, s_4, s_6\}$. Note that not all levels in the orderings contain worlds.

2. *$s \in P_n^t$ and for all $u \in S$, if $u \in P_x^s$ and $u \in P_y^t$ then $y \leq x$.*

One combination of analogs to the **S5** properties that may be of interest is:

- $\mathbf{K}'$: $(\nabla_i\phi \wedge \nabla_i(\phi \rightarrow \alpha)) \rightarrow \nabla_i\alpha$.
- $\mathbf{T}'$: $\nabla_{n-1}\phi \rightarrow \phi$.
- $\mathbf{4}''$: $\nabla_i\phi \rightarrow \nabla_{n-1}\nabla_i\phi$.
- $\mathbf{5}''$: $\neg\nabla_i\phi \rightarrow \nabla_{n-1}\neg\nabla_i\phi$.

These impose the following restrictions on our orderings:

- $\forall s \in S$, $s \notin P_n^s$
- $\forall s, t$ in S, if $t \notin P_n^s$ then $P^s = P^t$.

5. COMPARISONS WITH OTHER SYSTEMS

Kraus and Lehmann [14] combine the standard modal definitions of knowledge and belief and include in their system two modal operators on which are imposed the properties of **S5** and **KD45**. Formally, their system is comparable to what one might get from NOK by paying attention only to ∇_{n-1} and ∇_0. But their system does not ensure that any sensible or intuitive relationship exists between the two concepts, as each operator has its own accessibility relation. This makes it difficult to find any general properties relating the two constructs.

Halpern [11] defines a probabilistic account of knowledge and belief, based on the degree of certainty inherent in each notion. While this is similar to our approach in motivation, there is always the risk when following a numerical approach that one speaks in a spuriously precise manner about concepts or likelihoods that are far from exact. It is proposed in

that paper to instead work within a range of certainty values. NOK can accommodate such considerations by assigning to each world a probability or 'certainty' measure, and then ordering the worlds according to their relative measures.

Moses and Shoham [23] present a particularly interesting system in which belief is defined as defeasible knowledge. They set out with the intention to translate the phrase

> "The agent believes ϕ"

into

> "The agent knows that either ϕ is the case, or else some specific (perhaps unusual) circumstances obtain."

It turns out, however, to be necessary to make their belief operator *binary*; $B^{\alpha}\phi$ denotes ϕ is believed relative to the assumption α. In other words, $B^{\alpha}\phi$ means 'the agent knows that, assuming α holds, ϕ holds also'. The role of α is to exclude the unusual circumstances which might 'undercut' or 'defeat' ϕ. However, such an approach ignores the difficulties with exhaustive lists of exceptions familiar from 25 years of dealing with such problems as the qualification and frame problems [4].

NOK uses orderings of worlds, rather than frame axioms, to distinguish the normal from the exceptional. This makes NOK a system belonging to the tradition that can be traced back through Kraus, Lehmann and Magidor [15] to McCarthy [18, 19]. So in a way, it captures the intuitive idea behind Moses and Shoham's system without encountering the same difficulties.

We can view $\nabla_i\phi$ as stating

> "Assuming the state of the system is within those levels at or below P_i, then ϕ is true,"

and yet we need not explicitly denote what this assumption is, as it is inherent in the belief operator we are using.

Finally, van der Hoek and Meyer [27] introduce a system of graded modalities, $K_0, \ldots, K_n$, where $K_i\phi$ is true if and only if ϕ is *false* in at most i accessible states. The nature of this system is similar to NOK, in that agents are equipped with different degrees of certainty operators. It would seem possible to simulate this approach in NOK, without being limited to it.

NOK semantics attempts to draw together semantic information theory, epistemic logic, that tradition in nonmonotonic logic which we shall call preferential model semantics, and belief revision.

Preferential model semantics has its origin in the circumscription of McCarthy [19], the semantics of which involves an ordering on the interpretations of the (usually first-order) language. The approach was generalised and placed in a possible worlds framework by Shoham and Kraus, Lehmann, and Magidor [25, 24, 15, 16]. Given a set S of worlds and a suitable ordering (accessibility relation) $<$ on S, together with a function V that determines which atomic propositions are satisfied relative to which worlds, the idea is to broaden the usual semantic consequence relation $\models$ to a defeasible consequence relation, $\mid\!\sim$. Suppose $\mathcal{M}(\phi)$ is the set of worlds satisfying ϕ, and $\mathcal{M}in(\phi)$ the subset containing only those which are minimal in respect of the ordering $<$. Whereas $\phi \models \psi$ if and only if $\mathcal{M}(\phi) \subseteq \mathcal{M}(\psi)$, the extension of $\models$ to $\mid\!\sim$ is achieved by requiring instead that $\mathcal{M}in(\phi) \subseteq \mathcal{M}(\psi)$. In the case of infinite S, it is necessary to ensure that every set $\mathcal{M}(\phi)$ has minimal members. If S is taken to be the set of valuations of a finitely generated language, as in the preceding discussion, then every ordering on S is automatically well-founded.

NOK differs from preferential model semantics in associating a (potentially different) templated order with every world s. If the templated orderings are identical, then nevertheless NOK has a feature absent from preferential model semantics, namely the distinguished top level comprising the excluded worlds, which represents the marriage between semantic information theory and nonmonotonic logic.

Belief revision [10, 8] is of special relevance to NOK because of epistemic entrenchment — the idea that some beliefs may be more important to an agent than others. When forced to make a choice between two beliefs, the agent will discard the less entrenched belief.

There is a precise correspondence between epistemic entrenchment orderings and faithful total preorders on the set of states[22]. Specifically, every entrenchment ordering can be obtained from a total preorder on S by a suitable power construction.

A templated ordering assigns each world in S to one of $n+1$ different levels (P_0 to P_n), where n is the number of worlds in S. To translate this into a binary relation $\preceq$ on pairs of states, we may simply use the relative levels of each of the worlds: $a \preceq b$ iff $a \in P_i$ and $b \in P_j$ for some $i, j \in [0, n]$ with $i \leq j$. Expressing templated orderings as total preorders on S allows the connection between NOK and entrenchment to emerge.

THEOREM 3. *Let $\preceq$ be the total preorder associated with world $s \in S$, and $\sqsubseteq$ the corresponding entrenchment ordering. Then $\alpha \sqsubseteq \beta$ iff $\nabla_i\alpha \models \nabla_i\beta$ for all i.*

6. EXTENSIONS AND DIRECTIONS

Many connections require further exploration, including all those mentioned in the previous section, as well as the possibility of reconstructing NOK via a more conventional semantics involving an n-ary accessibility relation (a suggestion for which we thank our colleague Hans van Ditmarsch).

In the field of belief revision, exciting work is being done on the merging of orderings [17, 20]. Merging is at the heart of NOK and is likely to influence the obvious extension of NOK in which there is assigned to each world $s \in S$ not one but an array of templated orderings, each representing a different semantic notion (for example, risk).

A link that deserves a high priority is that with deontic logic and, more generally, the notion of trust, which is gaining importance from the development of agent communication languages and the need to enable verification of compliance [5].

Finally, the limitations of agents (such as resource-boundedness) are never absent when defeasible reasoning is the issue, and it will be interesting to see whether NOK lends itself to an investigation of such concerns despite the use of possible worlds. For example, it seems feasible to get around the logical omniscience problem by altering the semantics so that the statement $\nabla_i\phi$ is satisfied at some world s if ϕ is satisfied at *some sample* of the states in levels P_i^s and below. Under this setup, it is possible for $\nabla_i\phi$ and $\nabla_i(\phi \rightarrow \alpha)$ to hold for an agent without $\nabla_i\alpha$ being a necessary consequence. Exactly what the nature of the sample should be, and how we go about obtaining it, are questions that need answers.

7. REFERENCES

[1] Y. Bar-Hillel. An examination of information theory. *Philosophy of Science*, 22:86–105, 1955.

[2] Y. Bar-Hillel and R. Carnap. Semantic information. *British Journal for the Philosophy of Science*, 4:147–157, 1953.

[3] C. Boutilier. Unifying default reasoning and belief revision in a modal framework. *Artificial Intelligence*, 68:33–85, 1994.

[4] F. Brown. *The Frame Problem in Artificial Intelligence*. Morgan Kaufmann, Los Altos, CA, 1987.

[5] C. Castelfranchi and Y. Tan, editors. *Trust and Deception in Virtual Societies*. Kluwer, Dordrecht, in press.

[6] P. Cohen and H. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[7] E. Davis. *Representations of Commonsense Knowledge*. Morgan Kaufmann, 1990.

[8] P. Gärdenfors. *Knowledge in Flux: Modelling the Dynamics of Epistemic States*. The MIT Press, Cambridge, MA, 1988.

[9] P. Gärdenfors, editor. *Belief Revision*. Cambridge University Press, 1992.

[10] P. Gärdenfors and D. Makinson. Revisions of Knowledge Systems using Epistemic Entrenchment. In M. Vardi, editor, *Proceedings of the Second conference on Theoretical Aspects of Reasoning about Knowledge*, Los Altos, CA, 1988. Morgan Kaufmann.

[11] J. Halpern. The Relationship between Knowledge, Belief, and Certainty. *Annals of Mathematics and Artificial Intelligence*, 4:301–322, 1991.

[12] J. Hintikka and P. Suppes, editors. *Information and Inference*. D. Reidel, 1970.

[13] H. Katsuno and A. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52:263–294, 1991.

[14] S. Kraus and D. Lehmann. Knowledge, Belief and Time. *Theoretical Computer Science*, 58(1-3):155–174, June 1988.

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

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

[17] P. Maynard-Reid and Y. Shoham. Belief fusion: Aggregating pedigreed belief states. *Journal of Logic, Language, and Information*, 10(2):183–209, 2001.

[18] J. McCarthy. Epistemological Problems of Artificial Intelligence. In T. Kehler and S. Rosenschein, editors, *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pages 1038–1044, Los Altos, California, 1977. Morgan Kaufmann.

[19] J. McCarthy. Circumscription - A Form of Nonmonotonic Reasoning. In V. Lifschitz, editor, *Formalizing Common Sense: Papers by John McCarthy*, pages 142–157. Ablex Publishing Corporation, Norwood, New Jersey, 1980.

[20] T. Meyer. Merging epistemic states. In R. Mizoguchi and J. Slaney, editors, *Lecture Notes in Artificial Intelligence*, volume 1886, pages 286–296. Springer, 2000.

[21] T. Meyer, W. Labuschagne, and J. Heidema. Infobase change: a first approximation. *Journal of Logic, Language, and Information*, 9:353–377, 2000.

[22] T. Meyer, W. Labuschagne, and J. Heidema. Refined epistemic entrenchment. *Journal of Logic, Language, and Information*, 9:237–259, 2000.

[23] Y. Moses and Y. Shoham. Belief as Defeasible Knowledge. *Artificial Intelligence*, 64(2):299–321, 1993.

[24] Y. Shoham. A semantical approach to non-monotonic logics. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 388–392, 1987.

[25] Y. Shoham. *Reasoning about Change*. MIT Press, 1988.

[26] Y. Shoham. Agent-oriented programming. *Artificial Intelligence*, 60:51–92, 1993.

[27] W. van der Hoek and J.-J. Meyer. Graded Modalities in Epistemic Logic. In A. Nerode and M. Taitslin, editors, *Proceedings of the Logical Foundations of Computer Science*, pages 503–514. Springer, 1992.

APPENDIX

A. PROOFS OF THEOREMS

THEOREM 1. *Let $\mathcal{F}$ be a frame of NOK interpretations (S, F, V). Then $\mathcal{F} \Vdash \nabla_i \phi \rightarrow \nabla_{n-1} \nabla_i \phi$ iff for all $s, t \in S$ it is the case that if $t \notin P_n^s$ then either*

1. *$s \notin P_n^t$ and $P^t = P^s$, or*
2. *$s \in P_n^t$ and for all $u \in S$, if $u \in P_x^s$ and $u \in P_y^t$ then $x \leq y$.*

PROOF. ($\Leftarrow$) Take any NOK interpretation within $\mathcal{F}$. Further, take any world $s \in S$ and some sentence ϕ such that $\nabla_i \phi$ holds at s. We want to show $\nabla_{n-1} \nabla_i \phi$ also holds at s. Since s was arbitrary within S and our NOK-interpretation was arbitrary within $\mathcal{F}$, it then follows that $\mathcal{F} \Vdash \nabla_i \phi \rightarrow \nabla_{n-1} \nabla_i \phi$.

Let s' be some world not excluded by the ordering at world s, *i.e.,* some s' such that $s' \notin P_n^s$. Now, either condition (1) or (2) above must hold.

Assume (1) holds. That is, $s \notin P_n^{s'}$ and $P^{s'} = P^s$. Then, since $\nabla_i \phi$ holds at s, it must also hold at s'. Hence since s' was an arbitrary accessible world from s, $\nabla_{n-1} \nabla_i \phi$ must hold at s.

Assume instead that (2) holds. That is, $s \in P_n^{s'}$ and for all $u \in S$, we have if $u \in P_x^s$ and $u \in P_y^{s'}$, then $x \leq y$. Now consider the statement $\nabla_i \phi$ at world s'. For this to be false, there must be some world u in level P_i or below of the ordering at world s' at which ϕ does not hold. But this world u would have to be contained in some level P_x of the ordering at world s, with $i \geq x$. If such a world existed then $\nabla_i \phi$ would not hold at s. This is a contradiction. Hence no such world exists and so $\nabla_i \phi$ holds at s'. Again, since s' was an arbitrary accessible world from s, $\nabla_{n-1} \nabla_i \phi$ must hold at s.

($\Rightarrow$) Take $s, t \in S$ and assume $t \notin P_n^s$. Then we have two possibilities for the accessibility of world s from world t: accessible or inaccessible.

s **accessible from** t: Assume that s is accessible from world t, *i.e.*, $s \notin P_n^t$. Then we wish to show that $P^s = P^t$.

Assume the contrary, that $P^s \neq P^t$. Then there is some world u such that $u \in P_x^s$ and $u \in P_y^t$, with $x \neq y$.

Now, if $x > y$ then if we let ϕ be a voucher for all worlds in level P_y^s and below, *i.e.,* a sentence which is true in these worlds and these worlds *only*. Then $\nabla_y \phi$ holds at s but $\nabla_{n-1}\nabla_y \phi$ does not. This is because $u \in P_y^t$ and ϕ does not hold at u, so we have $\nabla_y \phi$ does not hold at t. Since t is accessible from s, this means $\nabla_{n-1}\nabla_y \phi$ cannot hold at s. But this contradicts our underlying assumption. Hence $x \not> y$.

Similarly, if $x < y$, then let ϕ be a voucher for the worlds in level P_{y-1}^t and below. Then $\nabla_{y-1}\phi$ holds at t, but $\nabla_{n-1}\nabla_{y-1}\phi$ does not. This is because $u \in P_x^s$ and ϕ does not hold at u, so we have $\nabla_{y-1}\phi$ does not hold at s. Since s is accessible from t, this means $\nabla_{n-1}\nabla_{y-1}\phi$ cannot hold at t. Again this is a contradiction. Hence we must conclude that $P^s = P^t$, as required.

s **inaccessible from** t: Assume that s is inaccessible from world t, *i.e.,* $s \in P_n^t$. Then we wish to show that for all $u \in S$, if $u \in P_x^s$ and $u \in P_y^t$, then $x \leq y$.

Assume the contrary, that there exists some $u \in S$ such that $u \in P_x^s$ and $u \in P_y^t$, with $y < x$. Then let ϕ be a voucher for the worlds in level P_{x-1}^s and below. Then $\nabla_{x-1}\phi$ holds at s, but $\nabla_{n-1}\nabla_{x-1}\phi$ does not. This is because $u \in P_y^t$ and ϕ does not hold at u, so we have $\nabla_{x-1}\phi$ does not hold at t. Since t is accessible from s, this means $\nabla_{n-1}\nabla_{x-1}\phi$ cannot hold at s. This too is a contradiction, so we must conclude that no such u exists. Hence for all $u \in S$, we have if $u \in P_x^s$ and $u \in P_y^t$, then $x \leq y$. □

THEOREM 2. *Let $\mathcal{F}$ be a frame of NOK interpretations (S, F, V). Then $\mathcal{F} \Vdash \neg\nabla_i\phi \rightarrow \nabla_{n-1}\neg\nabla_i\phi$ iff for all $s, t \in S$ it is the case that if $t \notin P_n^s$ then either*

1. *$s \notin P_n^t$ and $P^t = P^s$,* or
2. *$s \in P_n^t$ and for all $u \in S$, if $u \in P_x^s$ and $u \in P_y^t$ then $y \leq x$.*

PROOF. ($\Leftarrow$) Take any NOK interpretation within $\mathcal{F}$. Further, take any world $s \in S$ and some sentence ϕ such that $\neg\nabla_i\phi$ holds at s. We want to show $\nabla_{n-1}\neg\nabla_i\phi$ also holds at s. Let s' be some world which is accessible from s, so that $s' \notin P_n^s$.

Then as above, we have two possibilities for the accessibility of world s from world s': accessible or inaccessible.

s **accessible from** s'**:** Assume that s is accessible from world s', *i.e.,* $s \notin P_n^{s'}$. Then by condition (1) above we have $P^s = P^{s'}$. This means that since $\neg\nabla_i\phi$ holds at s, it must also hold at s'. Since s' was an arbitrary accessible world from s, we have $\nabla_{n-1}\neg\nabla_i\phi$ holding at s, as required.

s **inaccessible from** s'**:** Assume that s is inaccessible from world s', *i.e.,* $s \in P_n^{s'}$. Then by condition (2) above we have for all worlds $u \in S$, if $u \in P_x^s$ and $u \in P_y^t$ then $y \leq x$.

Now, since $\neg\nabla_i\phi$ holds at s, there must be some world w in level P_i^s or below where ϕ does not hold. Let k, j be the integers such that $w \in P_k^s$ and $w \in P_j^{s'}$. Then $j \leq k$ from (2). But $k \leq i$ from above, so we have $j \leq i$ also. Hence $\nabla_i\phi$ can't hold at world s' either. So $\neg\nabla_i\phi$ holds at s'. Again, since s' was an arbitrary accessible world from s, we have $\nabla_{n-1}\neg\nabla_i\phi$ holding at s, as required.

($\Rightarrow$) Let s, t be worlds in S with $t \notin P_n^s$. Then we have two possibilities for the accessibility of world s from world t: accessible or inaccessible.

s **accessible from** t: Assume that s is accessible from world t, i.e. $s \notin P_n^t$. Then we wish to show that $P^s = P^t$.

Assume the contrary, that $P^s \neq P^t$. Then there is some world u such that $u \in P_x^s$, $u \in P_y^t$, with $x \neq y$.

Now, if $x > y$ then if we let $\neg\phi$ be a voucher for world u, then $\neg\nabla_y\phi$ holds at t but $\nabla_{n-1}\neg\nabla_y\phi$ does not. This is because $\nabla_y\phi$ holds at world s, which is accessible from t, since $u \in P_x^s$ with $x > y$. So we have found a $y \leq n$ and a sentence ϕ such that $\neg\nabla_y\phi$ holds at t but $\nabla_{n-1}\neg\nabla_y\phi$ does not. But this contradicts our underlying assumption. Hence $x \not> y$.

Similarly, if $x < y$, then again let $\neg\phi$ be a voucher for world u, then $\neg\nabla_x\phi$ holds at s but $\nabla_{n-1}\neg\nabla_x\phi$ does not. This is because $\nabla_x\phi$ holds at world t, which is accessible from s, since $u \in P_y^t$ with $x < y$. So we have found a $x \leq n$ and a sentence ϕ such that $\neg\nabla_x\phi$ holds at s but $\nabla_{n-1}\neg\nabla_x\phi$ does not. Again, this is a contradiction. Hence $x \not< y$. So we must conclude that $P^s = P^t$, as required.

s **inaccessible from** t: Assume that s is inaccessible from world t, *i.e.,* $s \in P_n^t$. Then we wish to show that for all $u \in S$, if $u \in P_x^s$ and $u \in P_y^t$, then $y \leq x$.

Assume the contrary, that there exists some $u \in S$ such that $u \in P_x^s$, $u \in P_y^t$, with $y > x$. Then, as before, let $\neg\phi$ be a voucher for world u. Then $\neg\nabla_x\phi$ holds at world s but does not hold at world t. Hence it is not true at world s that $\neg\nabla_x\phi \rightarrow \nabla_{n-1}\neg\nabla_x\phi$. This is a contradiction. Hence no such u exists. Hence for all $u \in S$, if $u \in P_x^s$ and $u \in P_y^t$, then $y \leq x$. □

THEOREM 3. *Let $\preceq$ be the total preorder associated with world $s \in S$, and $\sqsubseteq$ the corresponding entrenchment ordering. Then $\alpha \sqsubseteq \beta$ iff $\nabla_i\alpha \models \nabla_i\beta$ for all i.*

PROOF. ($\Rightarrow$) Assume $\alpha \sqsubseteq \beta$. Let $A = M(\neg\alpha)$, $B = M(\neg\beta)$.

Take $b \in B$. Then if $\alpha \sqsubseteq \beta$ we have some $a \in A$ such that $a \preceq b$. Now, pick an element $b' \in B$ which is minimal in B, i.e. $\forall b'' \in B$, $b' \preceq b''$. Now, since $b' \in B$ and $\alpha \sqsubseteq \beta$, we must have some $a' \in A$ such that $a' \preceq b'$. This a' corresponds to a world where $\neg\alpha$ holds (since it is in A). Let t, u be the integers corresponding to the levels of a' and b' in the NOK ordering, i.e. t, u such that $a' \in P_t$ and $b' \in P_u$. Then because $a' \preceq b'$ we know that $t \leq u$.

Now, pick any q such that $0 \leq q \leq n$. Assume $\nabla_q\alpha$ holds. Then for all levels P_w, where $0 \leq w \leq q$, and for all worlds $s \in S$, if $s \in P_w$ then α holds at s.

Now, since $\neg\alpha$ holds at a', α cannot. So since $a' \in P_t$ we have $q < t$ (or else there is a world in level P_q or below for which α does not hold, and so $\nabla_q\alpha$ doesn't hold either — contradiction). Then, since $t \leq u$ and $q < t$, we have $q < u$.

Since P_u corresponded to the lowest level of the ordering which contained a world for which $\neg\beta$ held, and $q < u$, we must have all levels P_q and below holding only those worlds where β is true. So we get β holding in all worlds s, where $s \in P_w$ for some $w, 0 \leq w \leq q$. So $\nabla_q\beta$ holds.

Finally, since q was arbitrary, we have shown if $\alpha \sqsubseteq \beta$ is true then $\nabla_i\alpha \models \nabla_i\beta$ for all i.

($\Leftarrow$) Assume for all $i \leq n$ that $\nabla_i\alpha \models \nabla_i\beta$. We are required to show that $\alpha \sqsubseteq \beta$, i.e. if $A = M(\neg\alpha)$ and $B = M(\neg\beta)$, then for every $b \in B$ there is some $a \in A$ such that $a \preceq b$.

Let $A = M(\neg\alpha)$, $B = M(\neg\beta)$. Pick an arbitrary $b' \in B$. Assume there is no $a' \in A$ such that $a' \preceq b'$. Let q be the integer corresponding to the level of b' in the NOK ordering, i.e. q such that $b' \in P_q$. Then we have for all $a \in A$, if $a \in P_k$ then $k > q$. So no element of A resides in any level below or at level P_q in our NOK ordering. But this means that no world exists in level P_q or below in which α is false. Hence $\nabla_q \alpha$ holds.

But we just saw that β was false in world b', which is in level P_q. So $\nabla_q \beta$ does **not** hold. Thus we have found a $q \leq n$ such that $\nabla_q \alpha$ holds at s but $\nabla_q \beta$ does not. But this contradicts our initial assumption. Thus we must conclude that there is indeed some $a' \in A$ such that $a' \preceq b'$. Since b' was an arbitrary member of B, it follows that $\alpha \sqsubseteq \beta$.

Thus, if $\nabla_i \alpha \models \nabla_i \beta$ for all i, then $\alpha \sqsubseteq \beta$. $\square$