Information-theoretic 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

1. INTRODUCTION

The development of agent communication languages following the publication of [6, 30] has demonstrated the usefulness of epistemic logic as a core language to be enriched by additional operators, e.g. deontic operators or operators representing speech acts (see *e.g.* [9, 26]). While not disputing the value of the current emphasis on finding suitable explications of notions such as trust [5], we would suggest that more remains to be said about the core concepts of knowledge and belief. Drawing upon semantic information theory [2] and the idea that default rules, as realised by preference relations, give rise to information that is 'soft' or 'indefinite' [11], we explore a generalisation of the usual possible world semantics which, somewhat in the manner of Lewis's systems of spheres [19], assigns to every possible world a templated ordering of worlds.

In the following, we describe the information-theoretic origin of templated orderings, define our analog to Lewis's system of spheres, illustrate some of the possibilities for untangling customarily-conflated concepts by considering counterparts to the **KT45** properties, and compare the information-theoretic semantics with other approaches.

Before plunging ahead, we wish to make a point about generality. Our view of epistemic logic is agent-oriented rather than pure-mathematical. The agents in question are assumed to be concerned with representing knowledge about finite systems. Indeed, the systems will typically be 'diagrammable' in the sense that every component of the system can be named. For such systems a finitely generated propositional language suffices. We suggest that the class of epistemic languages which are of interest as precursors to practical agent communication languages are those having a finite set of atoms with a correspondingly finite set of valuations (of the underlying nonmodal language), and equipped with a semantics in which the set S of possible worlds is exactly the set of valuations. In the terminology of [25], we see no reason to augment the object states by meta-states. There is, in addition, a powerful metalogical reason for restricting attention to finitely generated languages. As shown below, in the case of a system whose description requires an infinite set of atomic propositions, there is semantic information that is ineffable (cannot be expressed syntactically). This would be far from desirable in an agent communication language.

2. SEMANTIC INFORMATION

Let S be the set of all valuations of some propositional language. The notion of information commonly 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]. Fortunately there is an alternative: the theory of semantic information introduced in [2] and extended by Hintikka (see *e.g.* [13]). It is remarkable that publications such as [8] make no reference to semantic information theory.

Semantic information theory takes the semantic content of a proposition ϕ ultimately to be determined by the valuations (possible worlds, members of S) that are "excluded" by ϕ, i.e. that are nonmodels of ϕ. 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 more fundamental than any ascription of content to a proposition.

THEOREM 1. *If S is infinite there exist $C \subseteq S$ such that for no proposition ϕ (or set Γ of propositions) is C the set of all models of ϕ (or of Γ).*

PROOF. Pick any valuation $s \in S$. Let $C = \{t : t \neq s\}$ and assume there exists a proposition α such that C is precisely the set of all models of α. Suppose that the atoms occurring in α are amongst $p_0, \dots, p_k$ and let s' be the valuation that coincides with s on $p_0, \dots, p_k$ while differing from s on all subsequent atoms. Then it is easy to see that if s' satisfies α (as it must), so does s, and thus s too must be a model of α. The argument for Γ is similar. □

The theorem shows that if S is infinite, an agent may have what we shall refer to as *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. We therefore take 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 $s \notin C'$. The theorem is not usually encountered in textbooks despite its elementary nature and deep significance; our attention was drawn to it by [4].

The theory of semantic information was devised for the case in which agents derive, from sensors or from other agents, information which is definite inasmuch as some set of possible worlds can be ruled out definitely. One thinks, for example, of a toe's impact against a rock as imparting definite information about the existence of an obstacle. However, agents may also acquire *indefinite* information via default rules. Whereas definite information has been represented semantically by a division of S into a complementary pair of subsets C and C', default rules may be represented by order relations on S [29, 7, 17]. To maintain unity between nonmonotonic logic and belief revision, the order relations on S that represent default rules may conveniently be taken to be either total preorders (as in the early belief revision literature, e.g. [14, 3]) or modular partial orders (e.g. [18, 22]), also called 'comparative relations' by van Benthem [31].

The essential feature of such orderings is that the set S is carved 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 observes the Light-Fan System. This system has just two components: a light and a fan. Each component is either on or off. Consider a default rule to the effect that the light and the fan are usually both on, less usually one off and the other on, and very seldom both off. Such a default rule could be expressed as an ordering that distributes the valuations representing the four possible states of the system among 3 layers. The bottom layer would hold the valuation according to which both components are on. The middle layer would hold those according to which only one component is on. The top layer would hold the valuation according to which neither component is on.

We propose to reconcile the notions of definite and indefinite information within a single data structure as follows.

Definition 1. Let the cardinality of S be n. The template for S is the chain $\{0, 1, \ldots, n\} = n + 1$ with the usual ordering. By a templated ordering of S we understand any function $f : S \longrightarrow n + 1$. The templated ordering f is in normal form if and only if, for each $i < n$, if $f^{-1}(i) = \varnothing$ then for all $i < j < n$, $f^{-1}(j) = \varnothing$ also. The templated ordering f is definite if and only if f is in normal form and for every $0 < i < n$, $f^{-1}(i) = \varnothing$; f is indefinite if and only if f is in normal form and $f^{-1}(n) = \varnothing$. The templated ordering f is contradictory if and only if $f^{-1}(n) = S$. The set of all templated orderings of S is denoted by T.

The idea is that the template for S possesses levels to some of which a templated ordering f assigns members of S, and the template is minimal in having only as many levels as might be required for a default rule that distinguishes between every pair of worlds while keeping separate a level dedicated to the definitely excluded worlds, should there be any. Note that no measure of distance is assumed, the intention being to represent qualitative comparison only.

While a templated ordering f is not an order relation, it induces one in the obvious way: $s \preceq t$ if and only if $f(s) \leq f(t)$. Clearly $\preceq$ is well-founded and a total preorder.

Definite information is represented by definite templated orderings — in these, members of S are assigned to either level 0 (the bottom level) or level n (the top level, representing the definitely excluded worlds in C'). Indefinite information is represented by a templated ordering assigning worlds in S to levels below n only, reflecting the intuition that a default rule does not suffice to definitely exclude any worlds. A templated ordering in normal form is one in which the worlds allocated to the levels below n are allocated without gaps (intervening empty levels); not all the levels below n will be nonempty, but the nonempty levels will be below all the empty levels. The top level is, however, the exception: even for a templated ordering in normal form it may be the case that level n is nonempty while one or more levels below it are empty. Note that a templated ordering in normal form will either have level 0 nonempty or be contradictory.

Level	Indefinite Info	Definite Info	Combination
4		00	00
3			
2	00		
1	10 01		10 01
0	11	11 10 01	11

Figure 1: 3 orderings reflecting indefinite information, definite information, and a combination of the two.

An agent would typically gain information about the system of interest by first acquiring definite information (either in the form of background information or by observation or by communicating with other agents) and then refining this with the help of its indefinite information. This may be represented by (renormalising the result of applying) the lexicographic refinement operation given by $f \boxplus g(s) = f(s)$ if $s \in C'$ else $g(s)$. Such operations are analogous to those explored in research on merging/fusion [21, 20]. See Figure 2. Notice that the top level of the templated ordering is crucial for the representation of definite information; it is unaffected by refinement, so that it is always possible to recover the agent's definite information.

3. INFORMATION-THEORETIC SEMANTICS

The usual possible worlds semantics provides for interpretations of the form (S, R, V) where S is an arbitrary set of states, R is an accessibility relation on S, and V is some function determining which atoms are satisfied at which states. Lewis generalises the possible worlds semantics by dropping the accessibility relation and instead assigning to every $s \in S$ a set of sets of states (subject to various restrictions). An information-theoretic semantics provides for interpretations in which every state s has a templated ordering assigned to it. (This resembles the update semantics of [15], except that those orderings represented similarity and lacked the top level whose information-theoretic importance has been demonstrated.)

Definition 2. Let L be the modal propositional language generated from the atoms $p_0, \ldots, p_m$ by the usual Boolean connectives and the modal operators $[0], \ldots, [n]$ where $n = 2^m$. Let S be the set of all valuations $s : \{p_0, \ldots, p_m\} \longrightarrow 2$. By an information-theoretic interpretation of L we understand a pair of the form (S, F) where $F : S \longrightarrow T^S$ is a function associating with each $s \in S$ a templated ordering of $F(s) \in T$.

Note that with loss of indefinite information an accessibility relation R on S may be recovered from each F. Simply take $R = \{(s, t) : F(s)(t) = 0\}$. Conversely, each R gives rise to an unique F. For each $s \in S$ let $F(s)$ be the definite templated ordering such that $F(s)(s') = 0$ if $(s, s') \in R$ else $F(s)(s') = n$.

Definition 3. Given an interpretation (S, F) and a world $s \in S$, we say that s satisfies the proposition α iff one of the following is the case:

- α is an atomic proposition and the valuation s satisfies α (*i.e.* $s(\alpha) = 1$)

- α is a Boolean combination of shorter propositions and is satisfied according to the usual truth-functional criteria
- $\alpha = [i]\beta$ and s' satisfies β for every s' such that $F(s)(s') \leq i$.

The idea is that $[i]\beta$ is satisfied at s if β is satisfied at all worlds s' lying at level i or below in the templated ordering associated with s. At one extreme, the operator $[0]$ represents the agent's most tentative defeasible beliefs, for $[0]\beta$ is satisfied at s if β is satisfied at the worlds which, according to the templated ordering $F(s)$, are most normal (occupy the bottom level). At the opposite extreme, $[n]$ is uninteresting and merely singles out the tautologies. However, $[n-1]$ is exceedingly interesting, since $[n-1]\beta$ is satisfied at s if β is satisfied at all the worlds that, according to the templated ordering $F(s)$, are not definitely excluded. We shall claim that the operators $[i]$ represent beliefs at various levels of entrenchment, with $[n-1]$ representing knowledge.

Epistemic entrenchment relations on propositions [10] are known to be closely related to total preorders on sets of worlds [22]. Specifically, every entrenchment relation can be obtained from a faithful total preorder on S by a suitable power construction. Recall that every templated ordering f on S is associated with a well-founded total preorder $\preceq$ on S. Well-foundedness coincides with faithfulness in our context, and so we have:

THEOREM 2. *Let $\preceq$ be the faithful total preorder on S associated with the templated ordering $F(s)$, and let $\sqsubseteq$ be the entrenchment ordering defined by requiring that $\alpha \sqsubseteq \beta$ if and only if for every nonmodel v of β there is a nonmodel u of α such that $u \preceq v$. Then $\alpha \sqsubseteq \beta$ if and only if s satisfies $[i]\alpha \to [i]\beta$ for all i (where $\to$ is the material conditional).*

PROOF. Note that the entrenchment ordering $\sqsubseteq$ is relative to s. For the slightly harder direction of the proof, assume that $[i]\alpha \to [i]\beta$ for all i but that it is not the case that $\alpha \sqsubseteq \beta$. Then there is some nonmodel v of β such that for every nonmodel u of α, $v \prec u$. Let $j = F(s)(v)$. Now s satisfies $[j]\alpha$ but not $[j]\beta$. $\square$

What is the significance of this result?

A tradition traceable back to Plato holds that knowledge is justified true belief [28]. There are difficulties surrounding justification [24] and these have led some to claim that knowledge is (merely) true belief [27]. Knowledge as true belief finds formal expression in the familiar axiom scheme T: $\Box\phi \to \phi$ (about which more later). However, the adjective 'true' indicates the adoption of a god's eye view rather than an agent-oriented view, agents rarely having complete information about the actual state of the world, even if they may be omniscient in the sense of having no resource bounds that limit knowledge of logical consequences. From an agent-oriented view, human experience suggests that knowledge is associated with an emotion that we may describe as 'reluctance to give up the belief', and such reluctance can clearly be modelled by epistemic entrenchment relations. In terms of entrenchment, an agent's most tentative beliefs occupy one end of a spectrum, and the most deeply entrenched beliefs the other end. This view needs qualification, as the most deeply entrenched beliefs are merely the tautologies. The theorem above describes an architecture of entrenchment in which, at each state s, there is information-theoretic justification for taking to be knowledge those beliefs that are entrenched to a specified level, namely the next-to-most deeply entrenched. These are the propositions α such that $\alpha \sqsubseteq \beta$ if and only if β is a tautology, and thus precisely the propositions α such that s satisfies $[n-1]\alpha$.

Taking the propositions α such that $[n-1]\alpha$ is satisfied at s to constitute the agent's knowledge at s, it would appear that the agent's beliefs form a hierarchy ranging from those reflecting knowledge to those reflecting tentative beliefs. This view is substantiated by the following:

THEOREM 3. *For all $j \leq n$, if s satisfies $[j]\alpha$ and $i < j$ then s satisfies $[i]\alpha$.*

The proof is trivial, but the significance is that if the agent believes α to a level j then the agent believes α to every lesser level (though not necessarily to any greater level).

4. CHARACTERISATIONS

There are many properties of templated ordering we may wish to impose. Of obvious interest is the class of agents which are rational, in the sense that they never believe contradictions. This corresponds to the class of interpretations (S, F) such that at every $s \in S$, the proposition $\neg[i]\mathit{falsum}$ is satisfied. These are precisely the interpretations such that for every $s \in S$, $F(s)^{-1}(0) \neq \varnothing$.

It is natural to consider the properties usually regarded as characterising knowledge, namely

- **K**: $(\Box\phi \wedge \Box(\phi \to \alpha)) \to \Box\alpha$
- **T**: $\Box\phi \to \phi$
- **4**: $\Box\phi \to \Box\Box\phi$
- **5**: $\neg\Box\phi \to \Box\neg\Box\phi$

Each of the above has a range of counterparts in information-theoretic epistemic logic. By way of illustration, we may mention

- $\mathbf{K}'$: $([i]\phi \wedge [i](\phi \to \alpha)) \to [i]\alpha$
- $\mathbf{T}'$: $[n-1]\phi \to \phi$
- $\mathbf{4}'$: $[i]\phi \to [n-1][i]\phi$
- $\mathbf{5}'$: $\neg[i]\phi \to [n-1]\neg[i]\phi$

$\mathbf{K}'$ asserts that if the beliefs ϕ and $\phi \to \alpha$ are entrenched to a certain level at s, so is α. All interpretations (S, F) satisfy this axiom scheme at all worlds in S.

$\mathbf{T}'$ asserts that if the agent knows ϕ (at s) then ϕ is true (at s). Interesting variants are possible, such as the scheme $[0]\phi \to \phi$ which asserts that even the most tentative beliefs of the agent are true.

$\mathbf{4}'$ asserts a powerful capacity for introspection, as the agent knows even its most tentative defeasible beliefs. Among the interesting variants would be the scheme $[n-1]\phi \to [0][n-1]\phi$ attesting to a weak form of introspection in which the agent is only dimly aware of its beliefs. Similarly with $\mathbf{5}'$ and negative introspection. Collectively the properties $\mathbf{K}'\mathbf{T}'\mathbf{4}'\mathbf{5}'$ characterise the class of agents in which 'mutually accessible' worlds have the same templated ordering associated with them.

THEOREM 4. *An interpretation (S, F) satisfies* $\boldsymbol{K'}\,\boldsymbol{T'}\,\boldsymbol{4'}\,\boldsymbol{5'}$ *at all worlds $s \in S$ if and only if for all $s, t \in S$ it is the case that*

1. $s \notin F(s)^{-1}(n)$, *and*
2. *if* $t \notin F(s)^{-1}(n)$ *then* $F(s) = F(t)$.

5. OTHER APPROACHES

The information-theoretic semantics described above represents one attempt to reconcile the notions of knowledge and belief and it seeks to accomplish its aim by building preference relations into the possible world semantics. The incorporation of preference as a general logical notion was suggested in [32] and explored from the perspective of dynamic logic. This paper changes the perspective to epistemic logic, and argues for a modified notion of preference, that of a templated ordering.

Good progress has been made in elucidating the relationship between information-theoretic semantics and the following approaches to reconciling knowledge and belief:

- Kraus and Lehmann [16] describe a system containing both a belief operator and a knowledge operator, satisfying the properties **KD45** and **KT45** respectively.
- Halpern [12] gives a probabilistic (*i.e.* numerical) account of knowledge and belief, based on the degree of certainty inherent in each notion.
- Van der Hoek and Meyer [33] introduce a fascinating system of graded modalities $[0], \ldots, [k]$ in which $[i]\phi$ is true at s if ϕ is false in at most i of the worlds accessible from s.
- Moses and Shoham [23] define various binary operators, representing defeasible belief, in terms of a standard knowledge operator.

Of special interest is the link with Lewis's systems of spheres. While the assignment of templated orderings is not a system of spheres, it induces in a natural way a system of spheres, and the reformulation of properties in the new context can be illuminating. For example, the requirement that agents not believe contradictions translates to the requirement that the systems of spheres be weakly centered.

It is our conviction that exploring these and other connections will show the information-theoretic semantics to be logically interesting and useful for applications.

6. 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] C. Brink and J. Heidema. A verisimilar ordering of propositional theories: the infinite case. Technical Report Technical Report 4/89, Department of Mathematics, Rand Afrikaans University, 1989.

[5] C. Castelfranchi and Y. Tan, editors. *Trust and Deception in Virtual Societies*. Kluwer, 2001.

[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] K. Devlin. *Logic and Information*. Cambridge University Press, 1991.

[9] FIPA. Communicative act library specification. Foundation for Intelligent Physical Agents, Document XC00037, http://www.fipa.org 1999.

[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] H. Geffner. *Default Reasoning: Causal and Conditional Theories*. MIT Press, 1992.

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

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

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

[15] H. Katsuno and A. Mendelzon. On the difference between updating a knowledge base and revising it. In P. Gärdenfors, editor, *Belief revision*. Cambridge University Press, 1992.

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

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

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

[19] D. Lewis. *Counterfactuals*. Basil Blackwell, 1973.

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

[21] T. Meyer. Merging epistemic states. In R. Mizoguchi and J. Slaney, editors, *Lecture Notes in Artificial Intelligence*, volume 1886, pages 286–296. Springer, 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] A. Musgrave. Rationalität und zuverlässigkeit. *Logos NF*, 7:94–114, 2001.

[25] M. Pagnucco and P. Peppas. Causality and minimal change demystified. In B. Nebel, editor, *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 125–130. Morgan Kaufmann, 2001.

[26] J. Pitt and Mamdani. Some remarks on the semantics

of fipa's agent communication language. *Autonomous Agents and Multi-Agent Systems*, 2(4):333–357, 1999.

[27] C. Sartwell. Why knowledge is merely true belief. *The Journal of Philosophy*, 89:167–180, 1992.

[28] F. Schmitt. *Knowledge and Belief*. Routledge, 1992.

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

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

[31] J. van Benthem. *The Logic of Time*. D. Reidel, 1983.

[32] J. van Benthem, J. van Eijck, and A. Frolova. Changing preferences. CWI report CS-R9310, 1993.

[33] 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.