The Logic of Epistemic Obligation and Two-dimensional Semantics

Xu Li *

University of Luxembourg, Luxembourg

xu.li@uni.lu

Within the broad realm of philosophical logic, deontic logic and epistemic logic are two fruitful branches. Compared to the effort devoted to each area, less attention has been paid to their interaction. Combing deontic and epistemic logic enables us to reason about notions such as "ought to know" and "ought to believe", which are crucial for practical issues like database security [3]. However, this problem is more difficult to solve than we thought, as suggested by Åqvist in his famous paradox of epistemic obligation [9]. Åqvist's paradox indicates that there exist apparent problems in a simple combination[1] of standard deontic logic (SDL or **KD**) and the very weak logic of knowledge **KT**. Consider the following scenario [7]:

(1) The bank is being robbed. (r)

(2) It ought to be the case that Jones (the guard) knows that the bank is being robbed. $(OK_j r)$

(3) It ought to be the case that the bank is not being robbed. $(O\neg r)$

Intuitively, we will agree that these statements describe a consistent scenario. However, if each sentence above is translated as the corresponding formula on the right, then the naive logic combining obligation and knowledge predicts that (1) – (3) form an inconsistent scenario, which contradicts our intuition. The reason is that $OK_j r \to Or$ is a theorem of the logic.

As suggested by Åqvist's paradox, any logical characterization of the notion "ought to know" should not have $OK_i\varphi \to O\varphi$ as a theorem. However, we note another closely related principle of "ought to know" that has been overlooked by almost all the literature. That is the principle (OKT): $OK_i\varphi \to \varphi$, stating that we have no obligation to know what is actually false.[2] According to the

*This work is supported by the Fonds National de la Recherche Luxembourg through the project Deontic Logic for Epistemic Rights (OPEN O20/14776480).

[1] By "simple combination", we mean to have a modal language with two (kinds of) modalities $K_i\varphi$ and $O\varphi$ for knowledge (of agent i) and obligation, respectively. The resulting logic **KD+KT** is obtained by combining all the axioms and rules of **KD** (for the modality O) and **KT** (for the modality K_i).

[2] As far as we know, this principle is only mentioned in the appendix of [1].

analysis of knowledge [6], most epistemologists have found it overwhelmingly plausible that what is false cannot be known. Given that "ought implies can", it is obvious that we have no obligation to know what is false. Thus, any logical formalization of "ought to know" should validate the principle (OKT). Clearly, (OKT) is not part of the previous logic **KD+KT**. However, if we assume a slightly stronger logic of knowledge, say **S4**,[3] and add (OKT) to **KD+S4**, unacceptable consequences follow:

(4) $OK_iK_i\varphi \to K_i\varphi$ (Instance of (OKT))

(5) $K_i\varphi \leftrightarrow K_iK_i\varphi$ (Axiom (T) and (4) of K)

(6) $OK_i\varphi \leftrightarrow OK_iK_i\varphi$ (Rule (RE) of O)[4]

(7) $OK_i\varphi \to K_i\varphi$ ((6), (4), propositional logic)

Though it is ideal for us to know what we ought to know, this is not always the case in reality. Note that we use only the rule (RE) of O and the axioms (T) and (4) of K. Thus, it seems that (OKT) can not be added, without invoking the problematic schema (7), into any minimal modal logic characterizing both obligation and knowledge in which knowledge is assumed to be veridical and positively introspective. This puts us in an awkward position.

Based on the previous analysis, we think that an adequate logical theory characterizing epistemic obligation should at least, first, avoid Åqvist's paradox and, second, validate the principle $OK_i\varphi \to \varphi$ (without invoking the problematic schema (7)). This paper aims to provide such a logical framework combining obligation, knowledge, and belief. The resulting logic generalizes the idea in Hilpinen's recent paper [5]. The core idea in [5] is the differentiation between wide scope and narrow scope interpretation of deontic modality. When we say "Jones ought to know (or to hold the true belief, if knowledge is analyzed as true belief) that the bank is being robbed", Hilpinen argues that the truth condition that the bank is being robbed is not in the scope of the deontic modality, but only Jone's belief is.

Following [5], the key component in our language is an auxiliary operator $A\varphi$ indicating that φ is not in the scope of the obligation modality O. Assume a countable infinite set of atoms PROP and a finite set of agents AG. The *language* $\mathcal{L}_{\{A,B,O\}}$ is generated by the following BNF grammar:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid A\varphi \mid B_i\varphi \mid O\varphi$$

where $p \in \textsc{prop}$ and $i \in \textsc{ag}$. Let $K_i\varphi$ and $P\varphi$ are abbreviations for $A\varphi \wedge B_i\varphi$ and $\neg O \neg\varphi$, respectively. $A\varphi$ reads as "Actually, φ is true" and $B_i\varphi$ "i believes that φ". Thus, $K_i\varphi$ is intended to express that "i knows φ". Last, $O\varphi$ expresses that "It is obligatory that φ".

A *model* is a tuple $M = (W, \{R_i\}_{i\in\textsc{ag}}, I, V)$ where W is a non-empty set; for each agent i, R_i is a binary relation on W; I is a serial binary relation

[3] Even if the modal logic **S5** is not accepted by some logicians as the right logic of knowledge, **S4** is generally accepted as a good representation of the properties of knowledge [8].

[4] (RE) of O: $\frac{\varphi\leftrightarrow\psi}{O\varphi\leftrightarrow O\psi}$ is a derived rule in **KD**.

on W; and $V : \text{PROP} \to \wp(W)$ is a valuation function. Given a model $M = (W, \{R_i\}_{i \in \text{AG}}, I, V)$, for every $w, w' \in W$ and formula φ, the *satisfaction relation* $\models \subseteq (W \times W) \times \mathcal{L}_{\{A,B,O\}}$ is inductively defined as follows (where the truth definition for Boolean connectives is as usual and, thus, omitted):

- $M, w, w' \models p$ iff $w' \in V(p)$
- $M, w, w' \models A\varphi$ iff $M, w, w \models \varphi$
- $M, w, w' \models B_i\varphi$ iff $M, v, v \models \varphi$ for all $v \in R_i(w')$
- $M, w, w' \models O\varphi$ iff $M, w', v \models \varphi$ for all $v \in I(w')$

A formula φ is *valid*, notation $\models \varphi$, if for all models M and $w \in M$, $M, w, w \models \varphi$.

The semantics we introduced is essentially two-dimensional [10]. The actuality operator $A\varphi$ has been interpreted using two-dimensional semantics in [2, 4], but in a way that one possible world in the model is designated as the actual. In our semantics, the formulas are evaluated with respect to a pair (w, w') of possible worlds (rather than a single world). w is called the "actual world" whereas w' is the "real" world for evaluating formulas. Normally, the actual world w and the evaluating world w' coincide ($w = w'$). Only within the scope of the O operator can they differ from each other. The semantics of $O\varphi$ keeps a record of the possible world at which $O\varphi$ is evaluated (the "actual world"). The information in the "actual world" can then be accessed by a formula like Ap within the (direct) scope of the O operator.

Example 1. The scenario described by the sentences (1) – (3) can be characterized by the following model M (we assume that Jones does not know that the bank is being robbed):

$$\begin{tikzcd} u : \neg r \arrow[loop above, dashed] & w : r \arrow[l, "j"] \arrow[r, dashed] & v : \neg r \arrow[loop above, dashed] \arrow[r, "j"] & x : r \arrow[loop above, dashed] \end{tikzcd}$$

where the doxastic accessibility relations are indicated by arrows with agents' names; the ideality relation is represented by dashed arrows; and, the current world is w. We have: (1) $M, w, w \models r$; (2) $M, w, w \models \neg K_j r$; (3) $M, w, w \models O\neg r$; and, (4) $M, w, w \models OK_j r$.

The above example shows that Åqvist's scenario can be represented in our framework. It can also be verified that the principle (OKT): $OK_i\varphi \to \varphi$ is valid, while the problematic schema $OK_i\varphi \to K_i\varphi$ does not follow in our framework (even when we have the **KD45**-logic for belief), as shown by the following counter model M.

$$\begin{tikzcd} u : \neg p \arrow[loop above, "i"] \arrow[loop below, dashed] & w : p \arrow[l, "i"] \arrow[r, dashed] & v : p \arrow[loop above, "i"] \arrow[loop below, dashed] \end{tikzcd}$$

Figure 1: A model M, where $M, w, w \not\models OK_i p \to K_i p$. Note that R_i is serial, transitive, and Euclidean.

References

[1] Guillaume Aucher, Guido Boella, and Leendert van der Torre. A dynamic logic for privacy compliance. *Artificial Intelligence and Law*, 19(2):187, 2011.

[2] John N Crossley and Lloyd Humberstone. The logic of "actually". *Reports on Mathematical Logic*, 8(1):1–29, 1977.

[3] Frédéric Cuppens and Robert Demolombe. A deontic logic for reasoning about confidentiality. In Mark A. Brown and José Carmo, editors, *Deontic Logic, Agency and Normative Systems*, pages 66–79, London, 1996. Springer London.

[4] Martin Davies and Lloyd Humberstone. Two notions of necessity. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 38(1):1–30, 1980.

[5] Risto Hilpinen. On the sick father, the repentant sinner, and other problems in medieval deontic logic. *Theoria*, 85(6):420–434, 2019.

[6] Jonathan Jenkins Ichikawa and Matthias Steup. The Analysis of Knowledge. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2018 edition, 2018.

[7] Paul McNamara and Frederik Van De Putte. Deontic Logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

[8] J.-J. Ch. Meyer and W. van der Hoek. *Epistemic Logic for AI and Computer Science*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, 1995.

[9] Lennart Åqvist. Good samaritans, contrary-to-duty imperatives, and epistemic obligations. *Noûs*, 1(4):361–379, 1967.

[10] Laura Schroeter. Two-Dimensional Semantics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.