*Logique & Analyse* **231** (2015), 395-415

# A SUITABLE SEMANTICS FOR IMPLICIT AND EXPLICIT BELIEF

ALESSANDRO GIORDANI

ABSTRACT

In the present paper a new semantic framework for modelling the distinction between implicit and explicit belief is proposed and contrasted with the currently standard framework based on the idea that explicit belief can be construed as implicit belief accompanied by awareness. It is argued that within this new framework it is possible to get both a more intuitive interpretation of the aforementioned distinction and a straightforward solution to two critical problems to which the standard view is subjected. A system of logic for belief is introduced and proved to be complete with respect to the class of all frames for implicit and explicit belief constructed in accord to the new view.

**Keywords**: awareness; epistemic logic; explicit belief; implicit belief; logical omniscience; possible worlds semantics.

## 1. Introduction

As is well-known, models of epistemic logic based on possible worlds semantics are subjected to the logical omniscience problem: epistemic agents believe all valid propositions and all the logical consequences of what they believe. This condition appears to be unsuitable for agents who, being characterized by epistemic limitations, such as lack of time, lack of power of deduction, lack of power of attention, or any combination thereof, are unable both to explicitly believe every proposition they could believe and to explicitly deduce all the consequences deriving from believed propositions.

The simplest solution to the omniscience problem consists in introducing a distinction between explicit and implicit belief and acknowledging that the set of implicit beliefs is closed under logical consequence, while the set of explicit belief is not[1]. In what follows, I will assess a now standard approach that allows us to avoid this problem along these lines and present

[1] Actually, there are different ways to try and model the distinction between explicit and implicit belief. See [20] ch. 1 for a general introduction and chs. 3 and 4 for a more in-depth analysis.

doi: 10.2143/LEA.231.0.3141817. 

a new semantics for systems of logic of implicit and explicit beliefs. The approach is the one proposed in [7], call it the *awareness approach*.[2] The basic idea is to model the distinction between explicit and implicit belief by using an awareness operator and construing explicit belief as implicit belief plus awareness. In addition, in this approach the basic intuition is preserved according to which the assumption that the set of implicit beliefs *is closed* under logical consequence is based on the fact that such set *includes* the closure under logical consequence of some smaller set, the set of explicit beliefs. A variant of the awareness approach[3], call it the *strong awareness approach*, is obtained if one makes the stronger assumption that the set of implicit beliefs *is identical with* the closure under logical consequence of the set of explicit beliefs, thus accounting for the fact that the notion of implicit belief seems to be derivative with respect to the notion of explicit belief.

The semantics based on the awareness approach is both simple and flexible: implicit belief is typically modelled on normal frames for epistemic logic as a *K45* or a *KD45* modality, whereas different conditions imposed on the set of propositions of which the agents are aware allow us to capture various interpretations of explicit belief.

In spite of its merits, this approach is not fully apt to furnish an intuitive understanding of the distinction between explicit and implicit belief, *if the set of implicit beliefs is construed as including the closure of the set of explicit beliefs*.

Let us assume both that a proposition is explicitly believed when it is actively held to be true by an agent and that any consequence of an explicit believed proposition is implicitly believed. Then, it can be shown that this intuitive understanding is not captured by the semantic model introduced to account for the behaviour of the awareness operator. Indeed, two central problems immediately arise. The first problem (*to which both the awareness approach and the strong awareness approach are subjected*) concerns the characterization of the explicitly believed propositions as the ones which are actively held to be true by an agent. Such a characterization is not seized by the definition of explicit belief as implicit belief plus awareness. Indeed, let *a* be an epistemic agent who actively held that the axioms of *PA2*[4], are

[2] The strong awareness approach can be implemented in different ways within diverse semantic frameworks. In particular, beside the semi-syntactic approach developed in [5], [7] and [8], a general semantic approach based on the addition of non standard impossible worlds to the standard possible ones has been proposed in [11] and developed in [15] and [21]. In this paper we will focus, without loss of generality, on the semantics for awareness as originated in [7]. The equivalence between this semantics and the possible / impossible worlds semantics was proved in [16].

[3] I want to thank an anonymous referee for pointing out to me the difference between the two approaches.

[4] Being finitely axiomatizable, *PA2*, i.e. second order Peano Arithmetic, can be explicitly held true.

true and φ be a theorem of *PA2*. Now, it is evident that *a* can both be aware of the content of φ and be uncertain about its truth: just imagine *a* trying to figure out if φ is provable within *PA2*. The second problem (*to which the strong awareness approach only is subjected*) concerns the characterization of the implicitly believed propositions as the ones which follow from what is explicitly believed. It will be shown that, within the semantic framework provided by the awareness approach, what is implicitly believed cannot be connected, as expected, with the set of propositions that are explicitly believed by an agent.

The paper is organized as follows. In Section 2 the possible worlds semantics for epistemic logic is briefly reviewed. In Section 3 the possible worlds semantics for awareness is introduced. In Section 4 this semantics is put into question. In Section 5 a new semantics is introduced in order to capture the fundamental intuition grounding the distinction between implicit and explicit belief as displayed above and a system of explicit logic is proved to be sound and complete with respect to this semantics. In Section 6 it is shown that the system of explicit logic is a conservative extension of the system *KD45* of belief. In section 7 a way to define the concept of awareness within the new semantics is scrutinized. Finally, in section 8 some possible developments are outlined.

In what follows, we will limit the discussion to the case where only belief is considered and one agent is involved. The generalization to the multi-agent case is straightforward.

## 2. Logic of belief[5]

Let $P$ be a set of propositional variables. The set $L(P, \boldsymbol{B})$ of epistemic formulas is inductively defined according to the following rules:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \varphi' \mid \boldsymbol{B}(\varphi)$$

where $p \in P$. The other propositional connectives are defined in the usual way. A frame for $L(P, \boldsymbol{B})$ is a pair $F = (W, R)$, where $W$ is a non-empty set of *worlds* and $R \subseteq W \times W$ is the *possibility* relation on $W$, modelling which worlds are to be considered possible from the point of view of any world $w$ in $W$. A model for $L(P, \boldsymbol{B})$ is a pair $M = (F, V)$, where $F$ is a frame for $L(P, \boldsymbol{B})$ and $V: P \rightarrow \wp(W)$ is a modal *valuation*, i.e. a function assigning to each propositional variable $p$ in $P$ a set of worlds in $W$. Intuitively $V$ assigns to each propositional variable $p$ the set of worlds in which $p$ is true.

[5] See [3] and [14] for an introduction to modal and epistemic logic. See [9] and [12] for detailed presentations.

**Definition 2.1**: $M,w \models \varphi$ ($\varphi$ is true at $w$ in $M$).

$M, w \models p \Leftrightarrow w \in V(p)$
$M, w \models \neg\varphi \Leftrightarrow$ not $M, w \models \varphi$
$M, w \models \varphi \wedge \varphi' \Leftrightarrow M, w \models \varphi$ and $M, w \models \varphi'$
$M, w \models \boldsymbol{B}(\varphi) \Leftrightarrow \forall v \in W(R(w, v) \Rightarrow M, v \models \varphi)$

Let $\varphi$ be a formula.
$\varphi$ is *valid* in $M$ iff it is true at every world in $M$ ($M \models \varphi$).
$\varphi$ is *valid* in a frame $F$ iff it is valid in every model based on $F$ ($F \Vdash \varphi$).
$\varphi$ is *valid* in a class of frames $\boldsymbol{F}$ iff it is valid in every frame in $\boldsymbol{F}$ ($\boldsymbol{F} \Vdash \varphi$).
We write $\Vdash \varphi$ to denote validity with respect to the class of all frames.
The foregoing semantics is subjected to the logical omniscience problem[6].
In particular, it is not difficult to show that each instance of the following schemas turns out to hold:

**LO1**: $\Vdash \varphi \Rightarrow \Vdash \boldsymbol{B}(\varphi)$
**LO2**: $\Vdash \varphi \rightarrow \varphi' \Rightarrow \Vdash \boldsymbol{B}(\varphi) \rightarrow \boldsymbol{B}(\varphi')$
**LO3**: $\Vdash \varphi \leftrightarrow \varphi' \Rightarrow \Vdash \boldsymbol{B}(\varphi) \leftrightarrow \boldsymbol{B}(\varphi')$
**LO4**: $\Vdash \boldsymbol{B}(\varphi \rightarrow \varphi') \rightarrow (\boldsymbol{B}(\varphi) \rightarrow \boldsymbol{B}(\varphi'))$

The concepts of *modal logic* and *normal modal logic* are the usual ones[7]. The basic normal modal logic $K$ is the smallest modal logic that contains all formulas of the form $\boldsymbol{B}(\varphi \rightarrow \varphi') \rightarrow (\boldsymbol{B}(\varphi) \rightarrow \boldsymbol{B}(\varphi'))$ and is closed under the necessitation rule: $\varphi / \boldsymbol{B}(\varphi)$. A normal modal logic generated by a set of axioms is the smallest normal modal logic that contains the axioms. The logics of belief considered here are normal modal logics in $L(P, \boldsymbol{B})$ generated by the following axioms:

| | | |
|---|---|---|
| *4*: | $\boldsymbol{B}(\varphi) \rightarrow \boldsymbol{B}(\boldsymbol{B}(\varphi))$ | positive introspection |
| *5*: | $\neg\boldsymbol{B}(\varphi) \rightarrow \boldsymbol{B}(\neg\boldsymbol{B}(\varphi))$ | negative introspection |
| *D*: | $\boldsymbol{B}(\varphi) \rightarrow \neg\boldsymbol{B}(\neg\varphi)$ | $\boldsymbol{B}$ consistency |

We denote with *KAx* the normal epistemic logic generated from $K$ by the list *Ax* of axioms on $\boldsymbol{B}$. Thus, *KD45* is the logic of belief that is typically considered to model the implicit belief of an *ideal* epistemic agent, which is intended to be an agent that does not implicitly believe contradictions. Let $\Lambda(\boldsymbol{F})$ be the set of formulas that are valid in the class of frames $\boldsymbol{F}$. It is straightforward to verify that $\Lambda(\boldsymbol{F})$ is a normal modal logic. Let $\Lambda$ be a normal modal logic and $\boldsymbol{F}(\Lambda)$ be the class of $F$ such that $F$ is a frame for $\Lambda$, where $F$ is said to be a *frame for* $\Lambda$ iff the logic of $F$ includes $\Lambda$.

[6] See [7], [8,ch.9] and [16] for analyses of both the problem and the principal strategies of solution.
[7] See [2], § 4.1.

**Definition 2.2**: soundness.
$\Lambda$ is said to be sound with respect to $\boldsymbol{F}$ iff $\Lambda \subseteq \Lambda(\boldsymbol{F})$ iff $\boldsymbol{F} \subseteq \boldsymbol{F}(\Lambda)$.

**Definition 2.3**: completeness.
$\Lambda$ is said to be complete with respect to $\boldsymbol{F}$ iff $\Lambda(\boldsymbol{F}) \subseteq \Lambda$.

**Definition 2.4**: characterization.
$\Lambda$ is said to characterize $\boldsymbol{F}$ iff $\boldsymbol{F}(\Lambda) = \boldsymbol{F}$.

An axiom is said to correspond to a condition on $R$ when the class of frames in which $R$ satisfies the condition is characterized by the logic generated by the axiom. It is well-known that any logic generated by a combination of axioms *D*, *4*, *5* is both complete and characterizes the class of frames generated by the combination of the corresponding (universally closed) conditions on $R$.

| logic | completeness | correspondence |
|---|---|---|
| *K* | all frames | no condition |
| *KD* | frames in which $R$ is serial | $\exists v R(w,v)$ |
| *K4* | frames in which $R$ transitive | $R(w,v)$ and $R(v,u) \Rightarrow R(w,u)$ |
| *K5* | frames in which $R$ Euclidean | $R(w,v)$ and $R(w,u) \Rightarrow R(v,u)$ |

## 3. Logic of explicit belief as a derived concept[8]

The intuition behind the awareness approach is that explicit belief implies implicit belief, whereas the converse does not hold in general. To explicitly believe an implicitly believed proposition an agent has to be aware of it, thus, in order to model awareness, a new modal operator, $\boldsymbol{A}$, is introduced. The set $L(P, \boldsymbol{B}, \boldsymbol{A})$ of formulas is then inductively defined according to the following rules:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \varphi' \mid \boldsymbol{B}(\varphi) \mid \boldsymbol{A}(\varphi)$$

A frame for $L(P, \boldsymbol{B}, \boldsymbol{A})$ is a tuple $F = (W, R, A)$, where $(W, R)$ is a frame for $L(P, \boldsymbol{B})$ and $A$ is a function that associates to each world $w$ in $W$ a set of formulas of $L(P, \boldsymbol{B}, \boldsymbol{A})$: intuitively, the set $A(w)$ of formulas that the agent is aware of at $w$. A model for $L(P, \boldsymbol{B}, \boldsymbol{A})$ is a pair $M = (F, V)$, where $F$ is a frame for $L(P, \boldsymbol{B}, \boldsymbol{A})$ and $V$ a valuation.

[8] See [7] and [8,ch.9] for detailed presentations.

**Definition 3.1**: $M,w \models \varphi$ ($\varphi$ is true at $w$ in $M$).
$M, w \models p \Leftrightarrow w \in V(p)$
$M, w \models \neg\varphi \Leftrightarrow$ not $M, w \models \varphi$
$M, w \models \varphi \wedge \varphi' \Leftrightarrow M, w \models \varphi$ and $M, w \models \varphi'$
$M, w \models \boldsymbol{B}(\varphi) \Leftrightarrow \forall v \in W(R(w, v) \Rightarrow M, v \models \varphi)$
$M, w \models \boldsymbol{A}(\varphi) \Leftrightarrow \varphi \in A(w)$

Within this framework explicit belief can be introduced by definition.

**Definition 3.2**: explicit belief (**b**).
$\mathbf{b}(\varphi) := \boldsymbol{B}(\varphi) \wedge \boldsymbol{A}(\varphi)$. Therefore, explicit belief = implicit belief + awareness.

**AL** (for awareness logic) is the system of normal modal logic for $\boldsymbol{B}$ generated by axioms *4* and *5*, and **IAL** (for ideal awareness logic) be the extension of **AL** obtained by adding *D*. Since no condition is imposed on $A$ and no axiom characterizes $\boldsymbol{A}$, system **AL** is sound and complete with respect to the class of all frames in which $R$ is transitive and Euclidean, while system **IAL** is sound and complete with respect to the class of all frames in which $R$ is serial, transitive and Euclidean. Furthermore, it is not difficult to see that the aforementioned omniscience principles turn out to be invalid once **b** is substituted for $\boldsymbol{B}$. (see [7]). Finally, notice that, since the logic of $\boldsymbol{B}$ coincides with *K45* and there is no interaction between $\boldsymbol{A}$ and $\boldsymbol{B}$, **AL** turns out to be a conservative extension of *K45*. Actually, any model of *K45* can be transformed into a model of **AL** satisfying the same set of $L(P,\boldsymbol{B})$ formulas by setting $A(w) = \varnothing$ for each $w$ in $W$. The same holds for **IAL** with respect to *KD45*.

Once a specific concept of awareness is at our disposal, the awareness function can be constrained in order to model it. As typical conditions, the following ones have been proposed (see, for instance, [10]):

| Conditions on $A$ | Corresponding axioms |
|---|---|
| *A1*: $\varphi\wedge\varphi' \in A(w) \Rightarrow \varphi\in A(w)$ and $\varphi'\in A(w)$ | **A1**: $\boldsymbol{A}(\varphi\wedge\varphi') \rightarrow \boldsymbol{A}(\varphi)\wedge \boldsymbol{A}(\varphi')$ |
| *A2*: $\neg\varphi\in A(w) \Rightarrow \varphi\in A(w)$ | **A2**: $\boldsymbol{A}(\neg\varphi) \rightarrow \boldsymbol{A}(\varphi)$ |
| *A3*: $\boldsymbol{A}(\varphi)\in A(w) \Rightarrow \varphi\in A(w)$ | **A3**: $\boldsymbol{A}(\boldsymbol{A}(\varphi)) \rightarrow \boldsymbol{A}(\varphi)$ |
| *A4*: $\boldsymbol{B}(\varphi)\in A(w) \Rightarrow \varphi\in A(w)$ | **A4**: $\boldsymbol{A}(\boldsymbol{B}(\varphi)) \rightarrow \boldsymbol{A}(\varphi)$ |
| *A5*: $\varphi\in A(w)$ and $\varphi'\in A(w) \Rightarrow \varphi\wedge\varphi'\in A(w)$ | **A5**: $\boldsymbol{A}(\varphi)\wedge \boldsymbol{A}(p') \rightarrow \boldsymbol{A}(\varphi\wedge\varphi')$ |
| *A6*: $\varphi\in A(w) \Rightarrow \neg\varphi\in A(w)$ | **A6**: $\boldsymbol{A}(\varphi) \rightarrow \boldsymbol{A}(\neg\varphi)$ |
| *A7*: $\varphi\in A(w) \Rightarrow \boldsymbol{A}(\varphi)\in A(w)$ | **A7**: $\boldsymbol{A}(\varphi) \rightarrow \boldsymbol{A}(\boldsymbol{A}(\varphi))$ |
| *A8*: $\varphi\in A(w) \Rightarrow \boldsymbol{B}(\varphi)\in A(w)$ | **A8**: $\boldsymbol{A}(\varphi) \rightarrow \boldsymbol{A}(\boldsymbol{B}(\varphi))$ |
| *A9*: $R(w,v) \Rightarrow A(w)\subseteq A(v)$ | **A9**: $\boldsymbol{A}(\varphi) \rightarrow \boldsymbol{B}(\boldsymbol{A}(\varphi))$ |
| *A10*: $R(w,v) \Rightarrow A(v) \subseteq A(w)$ | **A10**: $\neg\boldsymbol{A}(\varphi) \rightarrow \boldsymbol{B}(\neg\boldsymbol{A}(\varphi))$ |

Axioms **A1**-**A4** ensure that awareness is closed under taking sub-formulas. This is an intuitive but powerful property: if awareness is closed under sub-

formulas, then explicit belief is closed under implication[9]. Axioms **A1**-**A8** ensure that awareness is closed under taking all formulas generated by a certain set of propositional variables. This is a very strong property and it can be proved that, within a system of logic including **A1**-**A8**, awareness is definable in terms of explicit belief, since the equivalence $\boldsymbol{A}(p) \leftrightarrow \mathbf{b}(p) \vee \mathbf{b}(\neg\mathbf{b}(p))$ turns out to be valid[10]. Finally, a little thought shows that, given the other axioms, **A9** and **A10** ensure validity of the principles of positive and negative introspection for explicit belief.

## 4. Limits of the logic of explicit belief based on awareness

In what follows I think of the set of the epistemic states of an agent as a *database* consisting of information about both what propositions are taken into consideration and the way in which they are considered. The set containing the sentences expressing believed propositions can be called the *positive database*. In accordance with this model, explicit beliefs are beliefs concerning propositions in the positive database, while implicit beliefs are beliefs concerning propositions in the logical closure of that database.

It is now possible to display the limits of the awareness approach to the definition of the concept of explicit belief. The principal limit of this approach in modelling the intuitive concepts of implicit and explicit belief is given by the definition of explicit belief as implicit belief accompanied by awareness, where the set of implicit beliefs is identified with the set of the logical consequences of the set of the explicit beliefs[11]. Indeed, in the awareness approach, implicit belief and awareness are completely unrelated conditions: a glance at the truth-conditions for $\boldsymbol{B}$ and $\boldsymbol{A}$ suffices to show that implicit belief, in principle, has little to do with the propositions the agent is aware of, and thus with explicit belief. This consideration can be developed into two directions.

[9] Indeed, $\mathbf{b}(p \rightarrow p') \rightarrow (\mathbf{b}(p) \rightarrow (\boldsymbol{A}(p') \rightarrow \mathbf{b}(p')))$ is a valid formula of **IAL**. If awareness is closed under sub-formulas, then $\boldsymbol{A}(p \rightarrow p') \rightarrow \boldsymbol{A}(p')$; since $\mathbf{b}(p \rightarrow p') \rightarrow \boldsymbol{A}(p \rightarrow p')$, the conclusion follows.

[10] If $M, w \vDash \boldsymbol{A}(p)$, then $M, w \vDash \boldsymbol{A}(p) \wedge \boldsymbol{B}(p)$ or $M, w \vDash \boldsymbol{A}(p) \wedge \neg \boldsymbol{B}(p)$. In the first case, $M, w \vDash \mathbf{b}(p)$. In the other case, $M, w \vDash \boldsymbol{A}(p) \wedge \boldsymbol{B}(\neg \boldsymbol{B}(p))$, since $R$ is Euclidean, thus $M, w \vDash \boldsymbol{A}(p) \wedge \boldsymbol{B}(\neg \mathbf{b}(p))$, and, by **A6**-**A8**, $M, w \vDash \mathbf{b}(\neg \mathbf{b}(p))$. The other direction is straightforward, since awareness is closed under taking subformulas.

[11] This is the way in which the distinction is construed according to the standard awareness approach, where the set of implicit beliefs is introduced as a logically closed set containing the consequences of what is explicitly believed. See [8], pp. 337-8: "To represent the knowledge of agent $i$, we allow two modal operators, $K_i$ and $X_i$, standing for *implicit knowledge* and *explicit knowledge* of agent $i$, respectively. Implicit knowledge is the notion we have been considering up to now: truth in all worlds that the agent considers possible. On the other hand, an agent explicitly knows a formula $\varphi$ if he is aware of $\varphi$ and implicitly knows $\varphi$. Intuitively, an agent's implicit knowledge includes all the logical consequences of his explicit knowledge".

(I) *A problem for both the awareness approach and the strong awareness approach*.

Focusing on $\boldsymbol{A}$, it can be observed that agents are capable to understand, and thus be aware of, both believed and non-believed propositions. Still, if an agent can be aware of the proposition $p$ without being certain of its truth, it would be hardly intuitive to say that the agent explicitly believes $p$ only because it happens that $p$ is a consequence of something explicitly believed by her. In a similar sense, a rule like $\vdash p \Rightarrow \vdash \boldsymbol{A}(p) \rightarrow \mathbf{b}(p)$ is not valid with respect to human agents, even if idealized, since it excludes the possibility of looking for the solution of problems about logically true propositions, such as implications between axioms and theorems of a theory. In fact, in order to be engaged in a problem, an agent has both to be aware of the problem and to be uncertain about its solution. The same problem emerges when we consider the way in which an agent tries to make her implicit beliefs explicit. In this case, the agent can be aware of a proposition that follows from some premises she explicitly believes, and still fail to believe it explicitly, because she has to perform some inferential step in order to become aware of the truth of the proposition, and thus to explicitly believe it[12].

**Remark 1.** A different way to consider the same problem is to focus on the difference between explicit and implicit beliefs concerning contradictions. It is apparent that, once a proposition is discovered to imply a contradiction, it is rejected by any rational agent. Still, a contradictory proposition can be explicitly believed by an agent before being identified as such, as illustrated by Frege's Law V. Now, due to axiom $D$, within **IAL** no proposition can be both contradictory and explicitly believed. We can try to solve this problem by dropping $D$ and withdrawing to **AL**. Nevertheless, if the agent is aware of the possibility of a contradiction, this move provides no solution. Assume $\vdash_{\mathbf{AL}} \varphi \rightarrow \bot$, then

$\vdash_{\mathbf{AL}} \boldsymbol{B}(\varphi \rightarrow \bot)$, by the definition of $K$
$\vdash_{\mathbf{AL}} \boldsymbol{B}(\varphi) \rightarrow \boldsymbol{B}(\bot)$, by the definition of $K$
$\mathbf{b}(\varphi) \vdash_{\mathbf{AL}} \boldsymbol{B}(\varphi)$, by the definition of **b**
$\mathbf{b}(\varphi) \vdash_{\mathbf{AL}} \boldsymbol{B}(\bot)$, by logic
$\mathbf{b}(\varphi), \boldsymbol{A}(\bot), \vdash_{\mathbf{AL}} \mathbf{b}(\bot)$, by logic and the definition of **b**

As a consequence, we seem to be forced to conclude that, within **AL**, it is necessary for an epistemic agent to explicitly not believe the contradictory propositions the content of which she is aware. Frege could never have explicitly believed Law V.

[12] See [19] for an analysis of the case and the development of a model of epistemic dynamics. I want to thank an anonymous referee for drawing my attention to this case.

(II) *A further problem for the strong awareness approach*.

Focusing on $\boldsymbol{B}$, it can be observed that the concept of implicit belief is not a primitive one: implicit belief can be introduced as a modality characterizing what follows from propositions which are in the range of explicit belief (see [13], §1, [7], §3). In any case, provided that the set of implicitly believed propositions is a logically closed set and that the agent is incapable to explicitly believe all the propositions in this set, it seems to be necessary to refer to what is explicitly believed by the agent in order to understand what is determined as implicitly believed[13]. Still, this conceptual dependence is entirely missed within the awareness approach. In particular, every elementary proposition can be implicitly believed by an agent, irrespective of what the agent actually holds to be true. The following propositions show how this is possible.

**Proposition 4.1**: for every frame for $L(P, \boldsymbol{B}, \boldsymbol{A})$, every world $w$ in $W$, and every $p$ not in $A(w)$, there exists a model $M$ based on that frame such that $M, w \models \boldsymbol{B}(p)$.
Take $V$ such that $V(p) = \{v \mid R(w, v)\}$, for every $p \in P - A(w)$.

As a consequence, a proposition can be implicitly believed independently of its being a consequence of a set of explicitly believed propositions. In conclusion, the strong awareness approach is problematic because it assumes both (1) that explicit belief is definable as implicit belief plus awareness and (2) that implicit beliefs include propositions that logically follow from what is explicitly believed. A way out of this problem is simply to deny (2), developing an awareness approach in which what is implicit is not what follows from what is explicit (see e.g. [18]). Still, it is also of interest to stick to (2) and try to produce a suitable definition of the concept of implicit belief in terms of explicit belief and logical implication.

## 5. Logic of explicit belief as primitive concept

We can now pose the question whether it is possible, following our intuitions, both to introduce explicit belief as an independent concept and to model implicit belief as a modality characterizing the consequences of what is explicitly believed. To achieve our objective, an enrichment of the language is in order.

[13] In [4], §4, two other accounts of implicit belief are proposed. According to the first one, $\boldsymbol{B}(p)$ is to be construed as "$x$ believes or ought to believe that $p$". According to the other one $\boldsymbol{B}(p)$ is to be construed as "$x$ believes some proposition logically equivalent with $p$". In both cases, implicit belief is defined with reference to explicit belief.

The basic system of logic of implicit and explicit belief we are going to introduce is based on a language including a new propositional constant, $\mathsf{b}$, a new operator $\mathbf{b}$, and the global modality $\Box$[14]. The set $L(P, \mathsf{b}, \mathbf{b}, \Box)$ of formulas is then inductively defined according to the following rules:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \varphi' \mid \mathsf{b} \mid \mathbf{b}(\varphi) \mid \Box\varphi.$$

Intuitively, $\mathsf{b}$ is to be conceived as a constant referring to the content of the whole *positive database* of the agent. $\mathbf{b}$ is an operator that checks whether a proposition is contained in the database: $\mathbf{b}(\varphi)$ states that $\varphi$ is one of the explicitly believed propositions. Finally, $\Box$ is the global modality, checking whether a proposition is true at every world of a given model. $\Box$ is necessary in the present context in order to model the relation of logical consequence between propositions. In particular, $\Box$ is to be interpreted as stating that a proposition is logically true. It is worth noting that the relation of logical consequence is conceived here as a relation between propositions. Hence, $\Box(p_1 \rightarrow p_2)$ states that the proposition $p_2$ is, from the point of view of a model, a logical consequence of the proposition $p_1$, since $p_2$ is true at every world at which $p_1$ is true.

**Definition 5.1**: implicit ($\boldsymbol{B}$) and explicit ($\mathbf{b}$) belief.
i) $\varphi$ *is explicitly believed*: $\mathbf{b}(\varphi)$.
ii) $\varphi$ *is implicitly believed*: $\boldsymbol{B}(\varphi) := \Box(\mathsf{b} \rightarrow \varphi)$.

According to definition 5.1, a proposition is explicitly believed when it is a member of the agent's positive database, while it is implicitly believed when it follows, and in particular analytically follows, from explicitly believed propositions. Thus, definition 5.1 captures our basic intuitions. Let us consider now how to develop both a semantic framework and a system of logic for the previous notions.

A frame for $L(P, \mathsf{b}, \mathbf{b}, \Box)$ is a tuple $F = (W, B, C)$, where $B \subseteq W$ is a set of worlds, the ones that are possible from the point of view of the positive database, and $C$ is a function that associates to $\mathsf{b}$ a set of formulas, intuitively, the set $C(\mathsf{b})$ of formulas interpreted on propositions of the positive database. $C$ satisfies the *reflexivity condition*: $\mathsf{b} \in C(\mathsf{b})$. Finally, a model for $L(P, \mathsf{b}, \mathbf{b}, \Box)$ is a pair $M = (F, V)$, where $F$ is a frame for $L(P, \mathsf{b}, \mathbf{b}, \Box)$ and $V$ is a valuation that satisfy the subsequent *inclusion condition*.

**Definition 5.2**: $M, w \vDash \varphi$ ($\varphi$ is true at $w$ in $M$).

$M, w \vDash p \Leftrightarrow w \in V(p)$
$M, w \vDash \neg\varphi \Leftrightarrow$ not $M, w \vDash \varphi$

[14] See [2], ch.7, §1 for an introduction to this kind of modality.

$M, w \models \varphi \wedge \varphi' \Leftrightarrow M, w \models \varphi$ and $M, w \models \varphi'$
$M, w \models \mathsf{b} \Leftrightarrow w \in B$
$M, w \models \mathbf{b}(\varphi) \Leftrightarrow \varphi \in C(\mathsf{b})$
$M, w \models \Box\varphi \Leftrightarrow \forall v \in W(M, v \models \varphi)$

*Inclusion condition* (*IC*): for each $\varphi$, $w$, $\varphi \in C(\mathsf{b})$ and $w \in B \Rightarrow M, w \models \varphi$.

Let us note that $B$ is the set of worlds at which $\mathsf{b}$ is true, i.e. the set of worlds verifying the positive database of the agent. Thus, the set of all formulas that are true at every world in $B$ is precisely the set of what is implicitly believed given $\mathsf{b}$. $C(\mathsf{b})$ is a subset of this set, which is the subset of all formulas that are also explicitly believed: $\mathsf{b}$ is an element of $C(\mathsf{b})$, since $\mathsf{b}$ just refers to the content that she explicitly believes; in addition, any formula which is in $C(\mathsf{b})$ is explicitly believed by the agent. The inclusion condition is then essential for what is explicitly believed to play the correct function in determining what is implicitly believed. Its meaning is intuitive: each formula contained in the positive database is true at any world in which $\mathsf{b}$ is true, i.e. every formula that is explicitly believed is also implicitly believed. Let us also note that $\mathsf{b}$ represent the strongest proposition that is explicitly believed by the agent, since it implies any other explicitly believed proposition.

An important consequence of the foregoing definition is that the set of propositions that are explicitly believed by an agent is constant from a world to another. In fact, $B$ is the set of worlds that are possible from the point of view of the agent, and the point of view of such an agent is assumed to be characterized by her positive database $\mathsf{b}$. In addition, $\mathsf{b}$ is assumed to be the set of propositions that are actually and explicitly believed by the agent at a certain time. Hence, the identity of $\mathsf{b}$ is determined by the propositions it actually contains, which implies that the set of propositions contained in $\mathsf{b}$ is constant across possible worlds. This is why the function $C$, which specifies the content of $\mathsf{b}$, is introduced as a constant function[15].

Let **EL** (for *explicit belief logic*) be the basic system of normal modal logic for $\Box$ generated by the following axiom schemas.

Axioms and rules on $\Box$.

**A$\Box$1**: $\Box\varphi \rightarrow \varphi$.
**A$\Box$2**: $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$.
**R$\Box$**: $X \vdash \varphi \Rightarrow \Box X \vdash \Box\varphi$, where $X$ is a set of formulas and $\Box X = \{\Box\varphi \mid \varphi \in X\}$.

[15] It is possible to generalize this approach in order to make the database to vary from a world to another. Still, if we want both to do that and to be able to refer to the same database in different worlds, we are going to have to adopt a system of hybrid logic. This generalization has no effect on our central issue, and so we have preferred to adopt the simpler system proposed above.

Axioms on $\mathbf{b}$.

**Ab1**: $\mathbf{b}(\mathsf{b})$.
**Ab2**: $\mathbf{b}(\varphi) \rightarrow \Box\mathbf{b}(\varphi)$.
**Ab3**: $\mathbf{b}(\varphi) \rightarrow \Box(\mathsf{b} \rightarrow \varphi)$.

Notice that, given **A**$\Box$**1** and **Ab2**, **Ab3** is equivalent to $(\mathbf{b}(\varphi) \wedge \mathsf{b}) \rightarrow \varphi$.

**IEL** (for *ideal explicit belief logic*) is the extension of **EL** obtained by adding the axiom **Ab4**: $\neg\Box\neg\mathsf{b}$, stating that the positive database is consistent. It is worth noting that, if this database is consistent, then, even if it is possible to explicitly believe $\varphi$ without explicitly believing $\mathbf{b}(\varphi)$, it is not possible to explicitly believe $\varphi$ and to explicitly believe that $\neg\varphi$ is explicitly believed. Indeed, $\mathbf{b}(\varphi) \rightarrow \neg\mathbf{b}(\mathbf{b}(\neg\varphi))$ is a theorem of **IEL**:

$\vdash_{\mathbf{IEL}} \mathbf{b}(\varphi) \wedge \mathsf{b} \rightarrow \varphi$, by **Ab2** and **Ab3**
$\vdash_{\mathbf{IEL}} \mathbf{b}(\neg\varphi) \wedge \mathsf{b} \rightarrow \neg\varphi$, by **Ab2** and **Ab3**
$\vdash_{\mathbf{IEL}} \mathbf{b}(\varphi) \wedge \mathsf{b} \rightarrow \neg\mathbf{b}(\neg\varphi)$, by classical logic
$\vdash_{\mathbf{IEL}} \mathbf{b}(\mathbf{b}(\neg\varphi)) \wedge \mathsf{b} \rightarrow \mathbf{b}(\neg\varphi)$, by **Ab3**
$\vdash_{\mathbf{IEL}} \mathbf{b}(\varphi) \wedge \mathsf{b} \rightarrow \neg\mathbf{b}(\mathbf{b}(\neg\varphi))$, by classical logic
$\vdash_{\mathbf{IEL}} \Box\mathbf{b}(\varphi) \wedge \neg\Box\neg\mathsf{b} \rightarrow \neg\Box\mathbf{b}(\mathbf{b}(\neg\varphi))$, by **R**$\Box$ and classical logic
$\vdash_{\mathbf{IEL}} \Box\mathbf{b}(\varphi) \rightarrow \neg\Box\mathbf{b}(\mathbf{b}(\neg\varphi))$, by **Ab4**
$\vdash_{\mathbf{IEL}} \mathbf{b}(\varphi) \rightarrow \neg\mathbf{b}(\mathbf{b}(\neg\varphi))$, by **Ab2** and **A**$\Box$**1**

A similar derivation allows us to conclude: $\vdash_{\mathbf{IEL}} \mathbf{b}(\varphi) \rightarrow \neg\mathbf{b}(\neg\mathbf{b}(\varphi))$[16].

**Proposition 5.1**: **EL** is sound with respect to the class of all frames for $L(P, \mathsf{b}, \mathbf{b}, \Box)$.

*Proof*. The validity of the axioms of the first group is straightforward.
As to **Ab1**: $\mathsf{b} \in C(\mathsf{b})$, thus $\forall w(M, w \vDash \mathbf{b}(\mathsf{b}))$.
As to **Ab2**: let $M, w \vDash \mathbf{b}(\varphi)$; then $\varphi \in C(\mathsf{b})$ and $\forall w(M, w \vDash \mathbf{b}(\varphi))$.
As to **Ab3**: let $M, w \vDash \mathbf{b}(\varphi)$; then $\varphi \in C(\mathsf{b})$ and $\forall w(w \in B \Rightarrow M, w \vDash \varphi)$, by *IC*.

**Corollary**: **IEL** is sound with respect to the class of frames where $B$ is non-empty.
As to **Ab4**: assume $B \neq \varnothing$; thus $\exists v(v \in B)$; $\exists v(M, v \vDash \mathsf{b})$; $\forall w(M, w \vDash \neg\Box\neg\mathsf{b})$.

[16] As highlighted by an anonymous referee, this principle is interesting because of its connection with Moore's paradox. To be sure, if we assume that the set of explicit belief is closed under eliminating conjunction (see principle **b1** below), we get both $\vdash_{\mathbf{IEL}} \neg\mathbf{b}(\varphi \wedge \mathbf{b}(\neg\varphi))$ and $\vdash_{\mathbf{IEL}} \neg\mathbf{b}(\varphi \wedge \neg\mathbf{b}(\varphi))$. Thus, we are able to derive that both commissive and omissive Moorean sentences are not explicitly believable. In addition, in deriving the paradox we don't have to assume a principle of positive introspection for explicit belief, whereas the corresponding principle for implicit beliefs has to be assumed in the standard derivation of the paradox.

**Proposition 5.2**: **EL** is complete with respect to the class of all frames for $L(P, \mathsf{b}, \mathbf{b}, \Box)$.

*Proof*. Completeness is proved by canonicity: it is sufficient to prove that **EL** is complete with respect to the canonical model induced by a maximally **EL** consistent set.

Let $X$ be a **EL** consistent set of formulas. $X$ can be extended to a maximally consistent set $x$ by a standard procedure. In turn, $x$ induces a canonical model for **EL**. In what follows $w$, $v$, and so on, will range over maximally **EL** consistent sets.

**Definition 5.3**: canonical model induced by $x$.
Let $w/\Box = \{\varphi \mid \Box\varphi \in w\}$ and $w/\mathbf{b} = \{\varphi \mid \mathbf{b}(\varphi) \in w\}$.
The canonical model induced by $x$ is the tuple $\langle W_x, B_x, C_x, V_x\rangle$, where

– $W_x = \{w \mid x/\Box \subseteq w\}$
– $B_x = \{w \mid \mathsf{b} \in w \in W_x\}$
– $C_x$ is such that $C_x(\mathsf{b}) = x/\mathbf{b}$
– $V_x$ is such that $w \in V_x(p) \Leftrightarrow p \in w$

**Lemma 5.1**: $\langle W_x, B_x, C_x\rangle$ is a frame for **EL**.
1) $B_x \subseteq W_x$ by definition of $B_x$.
2) $\mathsf{b} \in C_x(\mathsf{b})$: $\mathbf{b}(\mathsf{b}) \in x$ by **Ab1**; hence, $\mathsf{b} \in x/\mathbf{b}$ and $\mathsf{b} \in C_x(\mathsf{b})$.

**Truth Lemma**: for every $\varphi$, $M_x, w \vDash \varphi \Leftrightarrow \varphi \in w$.
The only interesting cases are the ones concerning $\mathsf{b}$, $\mathbf{b}(\varphi)$ and $\Box\varphi$.

i) $M_x, w \vDash \mathsf{b} \Leftrightarrow w \in B_x \Leftrightarrow \mathsf{b} \in w$.
ii) $M_x, w \vDash \mathbf{b}(\varphi) \Leftrightarrow \varphi \in C_x(\mathsf{b}) \Leftrightarrow \varphi \in x/\mathbf{b} \Leftrightarrow \mathbf{b}(\varphi) \in x$.

By **Ab2**, $\mathbf{b}(\varphi) \in x \Leftrightarrow \mathbf{b}(\varphi) \in w$, whence the conclusion.

iii) $M_x, w \vDash \Box\varphi \Leftrightarrow \forall v(M_x, v \vDash \varphi) \Leftrightarrow \forall v \in W_x(\varphi \in v)$, by induction hypothesis.

By definition of $W_x$, $\cap \{v \mid v \in W_x\} = x/\Box$, since $x/\Box$ is closed with respect to derivability in **EL** and any **EL** closed set coincides with the intersection of the maximally **EL** consistent sets including it[17]. The conclusion follows.

**Lemma 5.2**: $\langle W_x, B_x, C_x, V_x\rangle$ satisfies *IC*.
Let $\varphi \in C_x(\mathsf{b})$ and $w \in B_x$, i.e. $\mathbf{b}(\varphi) \in x$ and $\mathsf{b} \in w$.
By **Ab2** we have $\Box\mathbf{b}(\varphi) \in x$; thus $\mathbf{b}(\varphi) \in w$.
By **Ab3** we have $\Box(\mathsf{b} \rightarrow \varphi) \in w$; thus $\varphi \in w$ and so $w \in V_x(\varphi)$, by the truth lemma.

[17] See [2], ch.4, for standard proofs of these facts.

**Corollary**: **IEL** is complete with respect to the class of frames where $B$ is non-empty.

We only have to check that $\langle W_x, B_x, C_x \rangle$ is a frame for **IEL**, i.e. that $B$ is non-empty. This follows from the fact that exists a $w$ such that $x/\Box \subseteq w$ and $\mathsf{b} \in w$. In turn, the existence of such a world follows from the consistency of $x/\Box \cup \{\mathsf{b}\}$, ensured by **Ab4**.

*Proof.* Assume that $x/\Box \cup \{\mathsf{b}\}$ is inconsistent. Then, there exist a finite set $X$ of formulas such that $\wedge(X) \in x/\Box$ and $\vdash_{\mathbf{IEL}} \wedge(X) \rightarrow \neg\mathsf{b}$, where $\wedge(X)$ denotes the conjunction of the formulas in $X$. Hence, $\vdash_{\mathbf{IEL}} \Box(\wedge(X) \rightarrow \neg\mathsf{b})$, $\vdash_{\mathbf{IEL}} \Box\wedge(X) \rightarrow \Box\neg\mathsf{b}$, and $\Box\neg\mathsf{b} \in x$, since $\Box\wedge(X) \in x$, contrary to the fact that $x$ is consistent and $\neg\Box\neg\mathsf{b} \in x$ by **Ab4**.

**Remark 2**. Systems **EL** and **IEL** can be extended by introducing axioms characterizing $\mathsf{b}$. The most intuitive axioms are the following ones.

**b1**: $\mathsf{b}(\varphi \wedge \varphi') \leftrightarrow \mathsf{b}(\varphi) \wedge \mathsf{b}(\varphi')$
**b2**: $\mathsf{b}(\mathsf{b}(\varphi)) \leftrightarrow \mathsf{b}(\varphi)$
**b3**: $\mathsf{b}(\Box\varphi) \rightarrow \mathsf{b}(\varphi)$
**b4**: $\mathsf{b}(\varphi) \rightarrow \neg\mathsf{b}(\neg\varphi)$
**b5**: $\mathsf{b}(\varphi \rightarrow \varphi') \wedge \mathsf{b}(\varphi) \rightarrow \mathsf{b}(\varphi')$
**b6**: $\mathsf{b}(\varphi \rightarrow \varphi') \wedge \mathsf{b}(\varphi' \rightarrow \varphi'') \rightarrow \mathsf{b}(\varphi \rightarrow \varphi'')$

It is not difficult to see that the axioms proposed below generate systems of explicit belief logic that are sound and complete with respect to classes of frames in which the function $C$ satisfies the following conditions.

*C1*: $\varphi \wedge \varphi' \in C(\mathsf{b}) \Leftrightarrow \varphi \in C(\mathsf{b})$ and $\varphi' \in C(\mathsf{b})$
*C2*: $\mathsf{b}(\varphi) \in C(\mathsf{b}) \Leftrightarrow \varphi \in C(\mathsf{b})$
*C3*: $\Box\varphi \in C(\mathsf{b}) \Rightarrow \varphi \in C(\mathsf{b})$
*C4*: $\varphi \in C(\mathsf{b}) \Rightarrow \neg\varphi \notin C(\mathsf{b})$
*C5*: $\varphi \rightarrow \varphi' \in C(\mathsf{b})$ and $\varphi \in C(\mathsf{b}) \Rightarrow \varphi' \in C(\mathsf{b})$
*C6*: $\varphi \rightarrow \varphi' \in C(\mathsf{b})$ and $\varphi' \rightarrow \varphi'' \in C(\mathsf{b}) \Rightarrow \varphi \rightarrow \varphi'' \in C(\mathsf{b})$

Let us call *self-reflecting* a database $\mathsf{b}$ that satisfies conditions *C1-C4* and *ideal self-reflecting* a database that satisfies conditions *C1-C6*.

A self-reflecting database is a database compiled by a very meticulous agent. To be sure: (i) if the agent writes down a conjunction, then she writes down both the conjuncts; vice versa, if she writes down both the conjuncts, then she writes down the conjunction; (ii) if the agent writes down that she explicitly believes that $\varphi$, then she writes down the witness of that belief, which is $\varphi$; in addition, if she writes down $\varphi$, then she explicitly believes that $\varphi$, and so she writes down that she explicitly believes that $\varphi$, which is $\mathsf{b}(\varphi)$; finally (iii) if the agent writes down that $\varphi$ is logically true, then she

writes down φ, and, if she writes down φ, then she doesn't write down that ¬φ, i.e. she is consistent in her explicit beliefs. Analogously, an ideal self-reflecting database captures the set of beliefs of an (somehow omniscient) agent whose explicit beliefs are closed under conditions *C1-C6*.

**Remark 3.** In **EL** what is believed, implicitly or explicitly, is logically determined, since a *specific positive database* is taken into consideration. In fact, it is easy to see that both $\Box\mathbf{b}(\varphi) \vee \Box\neg\mathbf{b}(\varphi)$ and $\Box\boldsymbol{B}(\varphi) \vee \Box\neg\boldsymbol{B}(\varphi)$ are valid formulas. The basic idea is that implicit and explicit beliefs are indexed with respect to $\mathsf{b}$, so that, once $\mathsf{b}$ is set, every conjunct in $\mathsf{b}$ and every consequence of $\mathsf{b}$ is determined at once. Therefore, saying that an agent believes a proposition is equivalent to saying that the agent believes a proposition in $\mathsf{b}$. In this sense, $\mathsf{b}$ can be interpreted as the database possessed by the agent at a certain time or context. As a consequence, $\mathbf{b}(\varphi)$ and $\Box\mathbf{b}(\varphi)$ have the same content, i.e. φ is explicitly believed in $\mathsf{b}$, so that $\Box\mathbf{b}(\varphi)$ says that it is necessary that φ is explicitly believed in $\mathsf{b}$, not that it is necessary that φ is explicitly believed. As a consequence, **EL** validates the principles of implicit positive and negative introspection with respect to both implicit and explicit beliefs:

$\vdash_{\mathbf{EL}} \mathbf{b}(\varphi) \rightarrow \boldsymbol{B}(\mathbf{b}(\varphi))$ $\quad$ $\vdash_{\mathbf{EL}} \boldsymbol{B}(\varphi) \rightarrow \boldsymbol{B}(\boldsymbol{B}(\varphi))$
$\vdash_{\mathbf{EL}} \neg\mathbf{b}(\varphi) \rightarrow \boldsymbol{B}(\neg\mathbf{b}(\varphi))$ $\quad$ $\vdash_{\mathbf{EL}} \neg\boldsymbol{B}(\varphi) \rightarrow \boldsymbol{B}(\neg\boldsymbol{B}(\varphi))$

By contrast, neither $\mathbf{b}(\varphi) \rightarrow \mathbf{b}(\mathbf{b}(\varphi))$ nor $\neg\mathbf{b}(\varphi) \rightarrow \mathbf{b}(\neg\mathbf{b}(\varphi))$ are derivable.

## 6. Relation between EL / IEL and *K45* / *KD45*.

In this section we present a proof that *KD45* can be embedded into **IEL** by first introducing a specific translation of the language of *KD45* into the language of **IEL** and then proving that a formula is derivable in *KD45* if and only if its translation is derivable in **IEL**. A proof that *K45* can be embedded into **EL** can be easily extracted from it. We will use the following translation of $L(P, \boldsymbol{B})$ into $L(P, \mathsf{b}, \mathbf{b}, \Box)$.

**Definition 6.1**: translation function *.
Let * be a the following translation of $L(P, \boldsymbol{B})$ into $L(P, \mathsf{b}, \mathbf{b}, \Box)$.
$p^* = p$
$(\neg\varphi)^* = \neg\varphi^*$
$(\varphi \wedge \varphi')^* = \varphi^* \wedge \varphi'^*$
$(\boldsymbol{B}(\varphi))^* = \Box(\mathsf{b} \rightarrow \varphi^*)$.

On this interpretation, and in accordance with definition 5.1, the implicit beliefs of the standard system of epistemic logic for belief are about the

propositions that follows from what is explicitly believed. Semantically, φ is implicitly believed just in case φ is true in every world in $B$, thus suggesting that $B$ corresponds to the set of world that are accessible from a reference world (a suggestion that will be exploited below in order to construct a **IEL** model from a *KD45* model).

**Proposition 6.1**: $X \vdash_{KD45} \varphi \Leftrightarrow X^* \vdash_{\mathbf{IEL}} \varphi^*$.
From left to right.
It suffices to show that the translations of *4*, *5*, *D* are theorems of **IEL**.

i) $\vdash_{\mathbf{IEL}} (\boldsymbol{B}(p) \rightarrow \boldsymbol{B}(\boldsymbol{B}(p)))^*$, i.e.
$\vdash_{\mathbf{IEL}} \Box(\mathsf{b} \rightarrow p) \rightarrow \Box(\mathsf{b} \rightarrow \Box(\mathsf{b} \rightarrow p))$.
$\Box(\mathsf{b} \rightarrow p) \vdash_{\mathbf{IEL}} \mathsf{b} \rightarrow \Box(\mathsf{b} \rightarrow p)$, by classical logic
$\Box(\mathsf{b} \rightarrow p) \vdash_{\mathbf{IEL}} \Box(\mathsf{b} \rightarrow \Box(\mathsf{b} \rightarrow p))$, by **R**□, **A**□**1**, **A**□**2**

ii) $\vdash_{\mathbf{IEL}} (\neg\boldsymbol{B}(p) \rightarrow \boldsymbol{B}(\neg\boldsymbol{B}(p)))^*$, i.e.
$\vdash_{\mathbf{IEL}} \neg\Box(\mathsf{b} \rightarrow p) \rightarrow \Box(\mathsf{b} \rightarrow \neg\Box(\mathsf{b} \rightarrow p))$.
$\neg\Box(\mathsf{b} \rightarrow p) \vdash_{\mathbf{IEL}} \mathsf{b} \rightarrow \neg\Box(\mathsf{b} \rightarrow p)$, by classical logic
$\neg\Box(\mathsf{b} \rightarrow p) \vdash_{\mathbf{IEL}} \Box(\mathsf{b} \rightarrow \neg\Box(\mathsf{b} \rightarrow p))$, by **R**□, **A**□**1**, **A**□**2**

iii) $\vdash_{\mathbf{IEL}} (\boldsymbol{B}(p) \rightarrow \neg\boldsymbol{B}(\neg p))^*$, i.e.
$\vdash_{\mathbf{IEL}} \Box(\mathsf{b} \rightarrow p) \rightarrow \neg\Box(\mathsf{b} \rightarrow \neg p)$.
$\mathsf{b} \wedge \Box(\mathsf{b} \rightarrow p) \wedge \Box(\mathsf{b} \rightarrow \neg p) \vdash_{\mathbf{IEL}} p \wedge \neg p$, by **A**□**1**
$\Box(\mathsf{b} \rightarrow p) \wedge \Box(\mathsf{b} \rightarrow \neg p) \vdash_{\mathbf{IEL}} \neg\mathsf{b}$, by classical logic
$\Box(\mathsf{b} \rightarrow p) \wedge \Box(\mathsf{b} \rightarrow \neg p) \vdash_{\mathbf{IEL}} \Box\neg\mathsf{b}$, by **R**□, **A**□**1**, **A**□**2**
$\Box(\mathsf{b} \rightarrow p) \vdash_{\mathbf{IEL}} \neg\Box(\mathsf{b} \rightarrow \neg p)$, by **Ab4** and classical logic

From right to left.

Since both systems are sound and complete, it suffices to show that any *KD45* model can be transformed into a **IEL** model validating the same formulas at any world.

Let $M = \langle W, R, V\rangle$ be a *KD45* and $x \in W$. We have to show that there is a **IEL** model $M^*$ and a world $x^*$ such that, for every φ, $M, x \vDash \varphi \Leftrightarrow M, x^* \vDash \varphi^*$.

Let $M^* = \langle W^*, B^*, C^*, V^*\rangle$, where

i) $W^* = \{x\} \cup \{w \mid R(x, w)\}$
ii) $B^* = \{w \mid R(x, w)\}$
iii) $C^*$ is defined by $C^*(\mathsf{b}) = \{\mathsf{b}\}$

Finally, $V^*$ is defined by $V^*(p) = V(p)$ and $V^*(\mathsf{b}) = B^*$.

**Lemma 6.1**: $M^*$ is an **IEL** model.

The only non-trivial point is to show that $B^*$ is non-empty.
$B^* = \{w \mid R(x, w)\}$; $R$ is serial; hence $\{w \mid R(x, w)\}$ is non-empty.

**Lemma 6.2**: for every $\varphi \in L(P, \boldsymbol{B})$, $M, x \vDash \varphi \Leftrightarrow M^*, x \vDash \varphi^*$.

The only interesting case is the modal one:

$M, x \vDash \boldsymbol{B}(\varphi) \Leftrightarrow \forall w(R(x,w) \Rightarrow M, w \vDash \varphi)$

$M, x \vDash \boldsymbol{BB}(\varphi) \Leftrightarrow \forall w(w \in B^* \Rightarrow M, w \vDash \varphi)$, by definition of $B^*$

$M, x \vDash \boldsymbol{B}(\varphi) \Leftrightarrow \forall w(w \in B^* \Rightarrow M^*, w \vDash \varphi^*)$, by induction hypothesis

$M, x \vDash \boldsymbol{B}(\varphi) \Leftrightarrow M^*, x \vDash \Box(\mathsf{b} \rightarrow \varphi^*)$, by definition of truth

$M, x \vDash \boldsymbol{B}(\varphi) \Leftrightarrow M^*, w \vDash (\boldsymbol{B}(\varphi))^*$, by definition of *

This ends the proof.

## 7. EL and awareness

In this last section we test **EL** with respect to the problems posed in §4. In particular, after defining a concept of awareness, we show that **EL** is not closed under the problematic rule $\vdash \varphi \Rightarrow \vdash \boldsymbol{A}(\varphi) \rightarrow \mathbf{b}(\varphi)$. As to the relation between explicit and implicit belief, it is plain from definition 5.1, ii), that a proposition is implicitly believed by an agent precisely when it is a consequence of her positive database.

In order to define awareness within our language, let us introduce a preliminary characterization of doubt, conceived as the state in which an agent is uncertain with respect to both a proposition and its negation. Let $\mathbf{?}(\varphi) := \neg\mathbf{b}(\varphi) \wedge \neg\mathbf{b}(\neg\varphi)$. $\mathbf{?}(\varphi)$ states that the positive database contains neither $\varphi$ nor $\neg\varphi$. This condition can obtain either because the agent is not aware of $\varphi$ or because the agent is aware of $\varphi$, but is not sure of its truth value. To capture the last condition we introduce the following definition.

**Definition 7.1**: doubt.

$\mathbf{?_b}(\varphi) := \mathbf{?}(\varphi) \wedge \mathbf{b}(\mathbf{?}(\varphi))$.[18]

According to definition 7.1, $\mathbf{?_b}(\varphi)$ states that $\varphi$ is problematic ($\mathbf{?}(\varphi)$) from the point of view of the agent ($\mathbf{b}(\mathbf{?}(\varphi))$), i.e. it is explicitly acknowledged that the positive database contains neither $\varphi$ nor $\neg\varphi$. Indeed, $\mathbf{?}(\varphi)$ describes a condition of explicit doubt about $\varphi$ and $\mathbf{b}(\mathbf{?}(\varphi))$ describes the explicit acknowledgement of that condition.

**Proposition 7.1**: given definition 6.2

1) $\mathbf{b}(\varphi) \rightarrow \neg\mathbf{?_b}(\varphi)$ is valid in **EL**: doubt is *excluded* by explicit belief.
2) $\boldsymbol{B}(\varphi) \rightarrow \neg\mathbf{?_b}(\varphi)$ is *not* valid in **IEL**: doubt is *not excluded* by implicit belief.

[18] In **EL** plus **b1**, **b2** and **b4** one can define $\mathbf{?_b}(\varphi)$ simply as $\mathbf{b}(\mathbf{?}(\varphi))$, since $\mathbf{b}(\mathbf{?}(\varphi)) \rightarrow \mathbf{?}(\varphi)$ is derivable.

*Proof*. 1): by definition of **?**(φ). 2): consider a model $M = \langle W, B, C, V \rangle$, where

i) $W = B = \{w_1, w_2\}$
ii) $C$ is such that $C(\mathsf{b}) = \{\mathsf{b}, \neg\mathbf{b}(p_1), \neg\mathbf{b}(\neg p_1)\}$
iii) $V$ is such that $V(p_1) = \{w_1, w_2\}$, else $V(p) = \varnothing$

It is evident that:

$M, w_1 \vDash \boldsymbol{B}(p_1)$, since $B = V(p_1)$.
$M, w_1 \vDash \mathbf{?_b}(p_1)$, since $p_1 \notin C(\mathsf{b})$, $\neg p_1 \notin C(\mathsf{b})$, $\neg\mathbf{b}(p_1) \in C(\mathsf{b})$, $\neg\mathbf{b}(\neg p_1) \in C(\mathsf{b})$.

At this point, it is possible to introduce awareness as the condition that characterizes a proposition φ if and only if φ is either believed or disbelieved or in doubt.

**Definition 7.2**: awareness.

$\boldsymbol{A}(\varphi) := \mathbf{b}(\varphi) \vee \mathbf{b}(\neg\varphi) \vee \mathbf{?_b}(\varphi)$ or, equivalently, $\mathbf{b}(\varphi) \vee \mathbf{b}(\neg\varphi) \vee \mathbf{b}(\mathbf{?}(\varphi))$.

Definition 7.2 is suitable in so far as it states conditions that are individually necessary and jointly sufficient for an agent to be aware of a proposition. To be sure, if an agent is in one of the above conditions, then she has to be aware of what she believes, disbelieves or doubts. On the other hand, if she is aware of a proposition, then it is reasonable to assume that the proposition has been classified according to one of the three basic epistemic modalities, i.e. believed, disbelieved, problematic.

**Proposition 7.2**: it follows from definition 7.2 that

1) $\boldsymbol{A}(\varphi) \wedge \boldsymbol{B}(\varphi) \rightarrow \mathbf{b}(\varphi)$ is *not* valid in **EL**.
2) $\vdash_{\mathbf{EL}} \varphi \Rightarrow \vdash_{\mathbf{EL}} \boldsymbol{A}(\varphi) \rightarrow \mathbf{b}(\varphi)$ is *not* a valid rule.
3) $\boldsymbol{A}(\varphi) \wedge \Box(\varphi' \rightarrow \varphi) \wedge \mathbf{b}(\varphi') \rightarrow \mathbf{b}(\varphi)$ is *not* valid in **EL**.
4) $\vdash_{\mathbf{EL}} \varphi' \rightarrow \varphi \Rightarrow \vdash_{\mathbf{EL}} \boldsymbol{A}(\varphi) \wedge \mathbf{b}(\varphi') \rightarrow \mathbf{b}(\varphi)$ is *not* a valid rule.

*Proof*. It is not difficult to see that 1) follows from 2) and 3) follows from 4).

2): consider a model $M$ in which $C(\mathsf{b}) = \{\mathsf{b}, \neg\mathbf{b}(p_1 \rightarrow p_1), \neg\mathbf{b}(\neg(p_1 \rightarrow p_1))\}$; then $\vdash_{\mathbf{EL}} p_1 \rightarrow p_1$ and, for each $w$, $M, w \vDash \mathbf{?_b}(p_1 \rightarrow p_1)$; thus, for each $w$, $M, w \vDash \boldsymbol{A}(p_1 \rightarrow p_1)$, but, for each $w$, not $M, w \vDash \mathbf{b}(p_1 \rightarrow p_1)$.

4): consider a model $M$ in which $C(\mathsf{b}) = \{\mathsf{b}, p_2, \neg\mathbf{b}(p_1 \rightarrow p_1), \neg\mathbf{b}(\neg(p_1 \rightarrow p_1))\}$; then $\vdash_{\mathbf{EL}} p_2 \rightarrow (p_1 \rightarrow p_1)$ and, for each $w$, $M, w \vDash \mathbf{?_b}(p_1 \rightarrow p_1)$ and $M, w \vDash \mathbf{b}(p_2)$; thus, for each $w$, $M, w \vDash \boldsymbol{A}(p_1 \rightarrow p_1) \wedge \mathbf{b}(p_2)$, but, for each $w$, not $M, w \vDash \mathbf{b}(p_1 \rightarrow p_1)$.

Notice that the same proof runs for every system between **EL** and **IEL, b1-b6**.

What about the problems proposed in section 4? We saw that (1) we should allow for an agent to be aware of a proposition $p$ without being certain of

its truth, even if $p$ follows from what the agent explicitly believes. Furthermore, (2) we should allow for an agent to be explicitly certain of a contradictory proposition, even if the agent is aware of a contradiction. Finally, (3) we should be able to prove that a proposition is not implicitly believed unless it is a consequence of a set of explicitly believed propositions. Ad (1): it is a straightforward consequence of 7.2 that $\boldsymbol{A}(\varphi) \wedge \Box(\varphi' \rightarrow \varphi) \wedge \mathbf{b}(\varphi') \wedge \neg\mathbf{b}(\varphi)$ is consistent. Hence, consistently with the example proposed in the introduction, an agent can explicitly believe a true proposition, be aware of one of its consequences, and still do not explicitly believe that very consequence. Ad (2): similarly, $\boldsymbol{A}(\bot) \wedge \Box(\varphi' \rightarrow \bot) \wedge \mathbf{b}(\varphi') \wedge \neg\mathbf{b}(\bot)$ is consistent. Hence, an agent can explicitly believe a contradictory proposition, be aware of a contradiction, and still do not explicitly believe a contradiction. Ad (3): by definition of implicit belief, every proposition that is implicitly believed follows from what the agent explicitly believe. It is worth noting that in a model in which $C(\mathsf{b}) = \{\mathsf{b}\}$ the agent can implicitly believe some atom $p_1$, but in this case the proposition corresponding to $p_1$ has to follow from the content of $\mathsf{b}$ (for example, $\mathsf{b}$ could correspond to $p_1 \wedge p_2 \wedge p_3$ under $V$, in the sense that $V(p_1) \cap V(p_2) \cap V(p_3) = B$). Note that we cannot fix a formula to which $\mathsf{b}$ is equivalent since the content of a formula changes when the valuation changes, while the content of $\mathsf{b}$ is constant.

**Remark 4**: The previous propositions are not surprising. Actually, even in **IEL** plus **b1-b6**, an agent is not necessarily *explicitly* certain of any instance of an axiom schema. This constraint can be weakened as follows. Let **A** be an axiom schema and $\mathbf{A_n} = \{\varphi \mid \varphi \text{ is an instance of } \mathbf{A} \text{ and } c(\varphi) = n\}$, where $c(\varphi)$ is the logical complexity of $\varphi$, determined by counting the number of logical constants in $\varphi$. Now, it is possible to impose that, for each $\varphi \in \mathbf{A_n}$, where $n$ is fixed and **A** is in a specified set of axiom schemas, if an agent is aware of $\varphi$, then $\mathbf{b}(\varphi)$, (take $C(\mathsf{b})$ such that $\varphi \in C(\mathsf{b})$ whenever $\varphi \in \mathbf{A_n}$ and $\{\neg\mathbf{b}(\varphi), \neg\mathbf{b}(\neg\varphi)\} \subseteq C(\mathsf{b})$).[19] However, even in this case, an agent can be aware of any $\varphi \in \mathbf{A_{n+1}}$ without being certain of its truth (take $C(\mathsf{b})$ such that $\varphi \notin C(\mathsf{b})$ and $\{\neg\mathbf{b}(\varphi), \neg\mathbf{b}(\neg\varphi)\} \subseteq C(\mathsf{b})$).

This shows that the present interpretation of $\boldsymbol{A}$, $\boldsymbol{B}$ and $\mathbf{b}$ is consistent with our intuitive judgements about the connections between awareness, implicit and explicit belief.

## 8. Further developments

This paper has introduced a basic logical system for implicit and explicit belief that provides a natural interpretation and coordination of these notions,

[19] This move gives us a certain control over an agent's computational resources.

since implicitly believed propositions are unambiguously identified with the consequences of the set of propositions that are held true by an agent. The system we have proposed can be developed in at least three directions. First, we can try to model implicit and explicit knowledge by introducing a constant $\mathsf{k}$, conceived as the conjunction of all propositions explicitly known by an agent, and a corresponding operator $\mathbf{k}$. Still, due to the presence of the global modality in the definition of implicit knowledge, this move can be successful only if we introduce an actuality operator, checking whether a proposition is true in the actual world[20]. Alternatively, we can introduce a knowledge operator by definition, stating that $\mathbf{k}(\varphi) := \mathbf{b}(\varphi) \wedge \varphi$[21]. If we want to do better, we have to abandon the reference to a special positive database and allow for the possibility for different possible worlds to be characterized by different positive databases. This is the second and more interesting line of development. The idea is to take the content of $\mathsf{b}$ as changing across worlds, so to model the general dynamics of explicit beliefs, and then to explore the possibility to define how $\mathsf{b}$ changes under a set of operations of deduction or observation, so to model the precise dynamics of explicit beliefs[22]. I also note that in such a context there are two ways in which the content of $\mathsf{b}$ can change across worlds: (i) the agent comes to believe a new independent proposition, so that both the set of explicit beliefs and the set of implicit beliefs changes; (ii) the agent comes to believe a dependent proposition, i.e. a proposition that follows from what she already believes, so that *only* the set of explicit beliefs changes. Finally, we can try to model implicit and explicit predicative knowledge along the same lines.

Alessandro GIORDANI
Department of Philosophy
Catholic University of Milan
L. A. Gemelli 1 – 20123 Milan
Italy
alessandro.giordani@unicatt.it

[20] To be sure, we have to specify (i) that $\mathsf{k}$ is true at the actual world and (ii) that both $\boldsymbol{K}(\varphi)$ and $\mathbf{k}(\varphi)$ imply that $\varphi$ is true at the actual world. The reflexivity of knowledge is not expressed by an axiom like $\Box(\mathsf{k} \rightarrow \varphi) \rightarrow \varphi$, because a consequence of such an axiom is $\Box(\mathsf{k} \rightarrow \varphi) \rightarrow \Box\varphi$, i.e. every known proposition is logically true. However, if we introduce the actuality operator @, then an axiom like $@\mathsf{k}$ provides a direct solution to (i) and an indirect solution to (ii), since the following implications are valid:

– $\mathbf{k}(\varphi) \rightarrow \Box(\mathsf{k} \rightarrow \varphi)$
– $\Box(\mathsf{k} \rightarrow \varphi) \rightarrow @(\mathsf{k} \rightarrow \varphi)$
– $@(\mathsf{k} \rightarrow \varphi) \rightarrow (@\mathsf{k} \rightarrow @\varphi)$

In this way we obtain reflexivity of knowledge in the form $\Box(\mathsf{k} \rightarrow \varphi) \rightarrow @\varphi$.

[21] I thank an anonymous referee for suggesting this further option.

[22] As suggested by an anonymous referee the approach proposed here could be usefully combined with the the ideas proposed in [18], ch. 5, in order to model inferential dynamics in general.

## References


[1] AGOTNES, T. and ALECHINA, N., The Dynamics of Syntactic Knowledge, Journal of Logic and Computation, 17(1), 2007: 83-116.

[2] BLACKBURN, P., M. DE RIJKE, and Y. VENEMA, Modal Logic. Cambridge: Cambridge University Press 2001.

[3] CHELLAS, B, Modal Logic: An Introduction, Cambridge: CUP 1980.

[4] CRESSWELL, J., Intensional Logics and Logical Truth, Journal of Philosophical Logic, 1, 1972: 2-15.

[5] CRESSWELL, J., Hyperintensional logic, Studia Logica 34, 1975: 25-38.

[6] H. N. DUC. Reasoning about rational, but not logically omniscient, agents, Journal of Logic and Computation, 7, 1997: 633–648.

[7] FAGIN, R. and HALPERN, J. Y., Belief, awareness, and limited reasoning, Artificial Intelligence, 34, 1988: 39-76.

[8] FAGIN, R., HALPERN, J., MOSES, Y., VARDI, M., Reasoning About Knowledge, Cambridge, MA: The MIT Press 1995.

[9] HALPERN, J.Y. and MOSES, Y., A guide to completeness and complexity for modal logics of knowledge and beliefs, Artificial Intelligence, 54, 1992: 319-379.

[10] HALPERN, J. Y., Alternative semantics for unawareness, Games and Economic Behaviour, 37, 2001: 321–339.

[11] HINTIKKA, J., Impossible Possible Worlds Vindicated, Journal of philosophical Logic, 4, 1975: 475-484.

[12] HOEK, W.v.d. and MEYER, J.J., Epistemic Logic for AI and Computer Science, Cambridge: CUP 1995.

[13] LEVESQUE, H.J., A logic of implicit and explicit belief, Proceedings AAAI-84, Austin, Texas, 1984: 198-202.

[14] MEYER, J.J., Epistemic Logic, Artificial Intelligence Preprint 1999.

[15] RANTALA, V., Impossible world semantics and logical omniscience. Acta Philosophica Fennica, 35, 1982: 106-115.

[16] THIJSSE, E. On total awareness logic. In de RIJKE, M. editor. Diamonds and Defaults, Berlin: Kluwer 1993.

[17] STALNAKER, R.C., Context and Content, Oxford: Oxford University Press 1999.

[18] VAN BENTHEM, J., Logical dynamics of information and interaction. Cambridge: CUP 2011.

[19] VAN BENTHEM, J. and VELÁZQUEZ-QUESADA, F.R., The dynamics of awareness, Synthese 117, 2010: 5–27.

[20] VELÁZQUEZ-QUESADA, F.R., Small steps in dynamics of information. Amsterdam: Institute for Logic, Language and Computation 2011.

[21] WANSING, H., A general possible worlds framework for reasoning about knowledge and belief, Studia Logica, 49, 1990: 523-539.