Avoiding logical omniscience by using subjective situations

Antonio Moreno[1], Ulises Cortés[2], Ton Sales[2]

[1] Departament d'Enginyeria Informàtica i Matemàtiques- Universitat Rovira i Virgili (URV)
Carretera de Salou, s/n. 43006-Tarragona, Spain - amoreno@etse.urv.es
[2] Dep. de Llenguatges i Sistemes Informàtics - Universitat Politècnica de Catalunya (UPC)
C/Jordi Girona, 1-3. 08034-Barcelona, Spain - {ia,sales}@lsi.upc.es

Abstract. The beliefs of the agents in a multi-agent system have been formally modelled in the last decades using *doxastic logics*. The *possible worlds model* and its associated *Kripke semantics* provide an intuitive semantics for these logics, but they commit us to model agents that are *logically omniscient*. We propose a way of avoiding this problem, using a new kind of entities called *subjective situations*. We define a new doxastic logic based on these entities and we show how the belief operators have some desirable properties, while avoiding logical omniscience. A comparison with two well-known proposals (Levesque's *logic of explicit and implicit beliefs* and Thijsse's *hybrid sieve systems*) is also provided.

1 Introduction

In the last decade *doxastic modal logics* have been considered the most appropriate formal tool for modelling the beliefs of the agents composing a multi-agent system ([1]). The standard way of providing a meaning to the modal formulas of these logics is to use the *possible worlds model* ([2]) and its associated *Kripke semantics* ([3]). This semantics is quite natural and intuitive, but it is well known that the agents modelled in this framework are *logically omniscient* ([4]). Therefore, this semantics is unsuitable to model the beliefs of realistic, non-ideal agents. The aim of our work is to provide a plausible way of modelling the beliefs of non-logically omniscient agents, while keeping the essence and the beauty of the possible worlds model and the Kripke semantics.

This article[3] is structured as follows. In section 2 we give an intuitive explanation of our approach to the logical omniscience problem, which is based in a new kind of entities called *subjective situations*. In a nutshell, a subjective situation is the perception that an agent has of a certain state of affairs. These situations, as will be explained below, will take the role of possible worlds. In section 3, a formalization of subjective situations in the framework of doxastic propositional logic is made. Section 4 is devoted to a study of the behaviour of the modal belief operators, that extends and generalizes our previous results ([5]). It is shown how their properties do indeed correspond with our intuitions about what should be an adequate formalization of the doxastic attitude of a non-ideal, non-logically omniscient agent. In section 5, a comparison of our proposal with two well-known approaches (Levesque's *logic of explicit and implicit beliefs* ([6])

[3] This research has been supported by the CICYT project *SMASH: Multi-agent systems and its application to hospital services (TIC96-1038-C04-04).*

and Thijsse's *hybrid sieve systems* ([7])) is performed. The paper finishes with a brief summary and the bibliographical references.

2 Motivation of subjective situations

The most popular way of dealing with the logical omniscience issue is to change the concept of what a *possible world* is (see [8] for a detailed review of the most interesting approaches to the problem of logical omniscience). Regardless of the way in which the concept of possible world is modified, there is a kernel that never changes: the formal representation of a possible world is not related in any way with the notion of *agent*. Thus, it may be said that all the approaches in the literature present an *objective* view of what a possible world is (*i.e.* a world is the same for all the agents, is independent of them). In a standard *Kripke structure*, the only item that depends on each agent is its *accessibility relation* between possible worlds.

The traditional meaning assigned to the accessibility relation R_i of an $Agent_i$ is that it represents the uncertainty that $Agent_i$ has about the situation in which it is located (*e.g.* $(w_0 R_5 w_1)$ means that $Agent_5$ cannot distinguish between worlds w_0 and w_1). This situation is quite peculiar, because the formulae that are true in two worlds that are linked by an accessibility relation are, in principle, totally unrelated (*i.e.* given a Kripke structure, there is no relationship between the accessibility relation between states and the function that assigns truth values to the basic propositions in each of them).

Our proposal may be motivated by the following scenario. Imagine two people (α and β) that are watching a football match together. In a certain play of the game, a fault is made and the referee awards a penalty kick. α thinks that the referee is right, because it has noticed that the fault was made inside the penalty area (let us represent this fact with proposition P); at the same time, β is thinking that the referee was wrong because, in its perception of the situation, the fault was made just an inch outside the penalty area. How can this situation (and the beliefs of the two agents) be formally represented?

Following the standard approach, we could model the fact that α believes P and β believes $\neg P$ by assuming that in all the (objectively described) worlds considered as possible in the current state by α the proposition P holds, whereas in all the worlds considered as possible by β (β's doxastic alternatives) P is false. This account of each agent's doxastic state does not seem very satisfactory to us, at least for two reasons:

- It does not tell us how each agent's perception of the situation influences in its own beliefs. An agent is supposed to eliminate instantly from its set of doxastic alternatives all those (completely specified) possible worlds in which a basic proposition has a truth value that does not match the agent's current beliefs. It would be more plausible to have a framework in which the agent kept a partial description of the situation in which it is located, and in which it could use the facts that it keeps perceiving from the environment in order to keep increasing and refining its beliefs ([9], [10]).
- Assuming that the fault was indeed made inside the penalty area, most philosophers would argue that α not only *believes* P but also *knows* it (being P true in the real world), whereas β believes $\neg P$ but can not possibly know $\neg P$, being it actually

false[4]. Thus, in a somehow *magical* way, one agent would have some knowledge (that would coincide with reality) whereas the other wouldn't.

In our opinion, this state of affairs (the actual situation, comprising both the football match and the agents, along with their beliefs) may not be adequately described with a simple assignment of truth values to the basic propositions. Even if we had an accurate description of the real world, does it really matter very much whether the fault was made inside the penalty area in order to model the beliefs of the two agents involved in the scene?

The situation (s) is obviously the same for the two agents α and β (they are watching the same match together). From α's point of view, the description of s should make true proposition P; however, from β's perspective, in the present situation P should be considered false. Obviously, there would be many aspects of s in which α and β would agree; *e.g.* both of them would consider that the proposition representing the fact *"We are watching a football match on TV"* is true in s.

As far as beliefs are concerned, we argue that, in this situation, α should be capable of stating that $B_\alpha P$ (α has seen the fault and has noticed that it was made inside the penalty area; thus, it believes so). It would not seem very acceptable a situation in which α perceived the fault to have been made inside the penalty area and defended that it did not believe that a penalty kick should have been awarded (the only possible explanation being that α is a strong supporter of the offending team). It also seems reasonable to say that α cannot fail to notice that it believes that the fault was made inside the penalty area; thus, α may also assert in s that $B_\alpha B_\alpha P$. In a similar way, in this situation β cannot state that $B_\beta P$ (β cannot defend that it believes that the referee is right, in a situation in which it perceived the fault to have been made outside the penalty area). Thus, it seems clear that each agent's point of view on a situation strongly influences (or we could say even *determines*) its positive and negative beliefs in that situation.

In our framework we want to include the intuition that agents are smart enough to know that other agents may not perceive reality in the same way as they do. In the previous example, without further information (*e.g.* α shouting *"Penalty!"*), β should not be capable of supporting (or rejecting) that $B_\alpha P$; analogously, α could not affirm (or deny) that $B_\beta P$. That means that the communication between the agents is the main way in which an agent may attain beliefs about other agent's beliefs. We could have chosen other alternatives; for instance, we could have stated that an agent believes that the other agents perceive reality in the same way as they do, provided that they do not have information that denies that fact. If that were the case α would assume that β also believes that P is true, as far as it does not have any reason not to think so (*e.g.* β saying *"This referee is really blind"*).

A final reflection on the meaning of the accessibility relation between situations for $Agent_i$ (R_i) is necessary. It will be assumed that an agent cannot have any doubts about its own perceptions and beliefs in a given state. *E.g.* if, in situation s, α looks at the match and thinks P, then it surely must realise this fact and believe P in s (and even

[4] It could be argued that we are somehow neglecting the need of a *justification* for the belief in order for it to become knowledge (as knowledge is usually defined in the philosophical literature as *true justified belief*). But, what could possibly count more as a justification that each agent's own direct perception of the situation?

believe that it believes P, were it to think about that). Thus, if R_α links s with all those situations that α cannot tell apart from s, it must be the case that α also perceives P as true in all those states as well (otherwise, those states would be clearly distinguishable by α, because in some of them it would support P whereas in some of them P would be rejected). The only uncertainty that α may have is *about the perception of s by the other agents*. In the example, α does not know whether it is in a situation in which β supports P or in a situation in which β rejects P. Therefore, α's accessibility relation must reflect this uncertainty.

Summarising, the main points that have been illustrated with the previous discussion are the following:

- A situation may be considered not as an entity that may be objectively described, but as a piece of reality that may be perceived in different ways by different agents. Thus, it is necessary to think of a *subjective* way of representing each situation, in which each agent's point of view is taken into account. In the previous example, the description of s should include the fact that α is willing to support P, whereas β isn't.
- An agent's beliefs in each situation also depend on its point of view.
  In the situation of the example, $B_\alpha P$ would hold from α's perspective, whereas it would not be either supported or rejected by β. Thus, we argue that it does not make sense to ask whether $B_\alpha P$ holds in s or not; that question must be referred *to a particular agent's point of view*.
- The interpretation of the meaning of each agent's accessibility relation is slightly different from the usual one.
  Each accessibility relation R_i will keep its traditional meaning, *i.e.* it will represent the uncertainty of $Agent_i$ with respect to the situation in which it is located. However, our intuition is that an agent may only be uncertain about the other agents' perception of the present state, not about its own perception.

3 Formalization of subjective situations

These intuitive ideas are formalized in the *structures of subjective situations*:

Definition 1 (Structure of Subjective Situations)
An structure of subjective situations for n agents is a tuple

$$< S, R_1, ..., R_n, \mathcal{T}_1, ...\mathcal{T}_n, \mathcal{F}_1, ..., \mathcal{F}_n >, \textit{ where}$$

- *S is the set of possible situations.*
- *R_i is the accessibility relation between situations for $Agent_i$.*
- *$\mathcal{T}_i$ is a function that returns, for each situation s, the set of propositional formulae that are perceived as true by $Agent_i$ in s.*
- *$\mathcal{F}_i$ is a function that returns, for each situation s, the set of propositional formulae that are perceived as false by $Agent_i$ in s.*

$\mathcal{E}$ is the set of all structures of subjective situations.

The presence of $\mathcal{T}_i$ and $\mathcal{F}_i$ allows $Agent_i$ to consider *partial* situations (those in which $Agent_i$ does not have any reason to support or to reject a given formula) as well as *inconsistent* situations (those in which $Agent_i$ may have reasons to support and to reject a given formula). This kind of situations was already considered by Levesque in his *logic of explicit and implicit beliefs* ([6]). A detailed comparison of our proposal and that of Levesque is offered in section 5.

The accessibility relation between situations for $Agent_i$ has to reflect its uncertainty about the way in which the actual situation is perceived by the other agents. Thus, R_i has to link all those states that $Agent_i$ perceives in the same way but that may be perceived in different ways by other agents. This intuition is formalized in the following condition:

Definition 2 (Condition on Accessibility Relations)

$$\forall s,t\epsilon S, (sR_it) \textit{ if and only if } (\mathcal{T}_i(s) = \mathcal{T}_i(t)) \textit{ and } (\mathcal{F}_i(s) = \mathcal{F}_i(t))$$

This condition implies that the accessibility relations are equivalence relations. This result links this approach with the classical $S5$ modal system, in which this condition also holds. In $S5$ the presence of this condition makes true axiom *4* (positive introspection), axiom *5* (negative introspection) and axiom *T* (the axiom of knowledge); the modal operators of the system proposed in this article will have similar properties, as will be shown in section 4.

3.1 Satisfiability relations

A simplified version of the doxastic propositional language for n agents is considered, as shown in the following definition:

Definition 3 (Doxastic Modal Language $\mathcal{L}$)
Consider a set of modal belief operators for n agents (B_1, ..., B_n). $\mathcal{L}$ is the language formed by all propositional formulae (built in the standard way from a set $\mathcal{P}$ of basic propositions and the logical operators $\neg, \vee, \wedge, \rightarrow$), preceded by a (possibly empty) sequence of (possibly negated) modal operators. $\mathcal{L}_{PC}$ is the subset of $\mathcal{L}$ that contains those formulae that do not have any modal operator. The modal formulae of $\mathcal{L}$ are called linearly nested*.*

Thus, the language $\mathcal{L}$ contains formulae such as P, B_3Q, $B_1B_5(R \vee T)$, $B_3\neg B_2S$ and $\neg B_1B_1\neg T$, but it is not expressive enough to represent formulae such as $(B_2P \rightarrow B_3Q)$ or $(P \vee B_5Q)$. In most practical applications, an agent in a multi-agent system will only need to represent what it believes (or not) to be the case in the world and what it believes (or not) that the other agents believe (or not). This is just the level of complexity offered by *linearly nested* formulae.

In an structure of subjective situations each $Agent_i$ may have *positive* and *negative* information about some propositional formulae (given by $\mathcal{T}_i$ and $\mathcal{F}_i$, respectively). This allows us to define two relations (of satisfiability, $\models_i$, and unsatisfiability, $\dashv_i$) between situations and formulae *for each* $Agent_i$. Given an structure of subjective situations E and a situation s, the expression $E, s \models_i \phi$ should hold whenever $Agent_i$ has some reason to think that ϕ is true in situation s. Similarly, $E, s \dashv_i \phi$ should hold whenever $Agent_i$ has some reason to reject ϕ in situation s.

Notice that $E, s \not\models_i \phi$ should not imply that $E, s \dashv_i \phi$ (i.e. $Agent_i$ not having any reason to support ϕ does not mean that it must have reasons to reject it). In the same spirit, $E, s \models_i \phi$ should not imply that $E, s \not\dashv_i \phi$ ($Agent_i$ could have reasons both to support and to reject a certain formula in a given situation). These facts will indeed be true, as will be seen in the next section, due to the presence of partial and inconsistent situations commented above.

The clauses that define the behaviour of these relations are shown in the following definition:

Definition 4 (Relations $\models_i$ and $\dashv_i$)

– $\forall E \epsilon \mathcal{E}, \forall s \epsilon S, \forall agent\, i, \forall \phi \epsilon \mathcal{L}_{PC}$

$$E, s \models_i \phi \Leftrightarrow \phi \epsilon \mathcal{T}_i(s)$$
$$E, s \dashv_i \phi \Leftrightarrow \phi \epsilon \mathcal{F}_i(s)$$

– $\forall E \epsilon \mathcal{E}, \forall s \epsilon S, \forall agents\, i, j,\ \forall \phi \epsilon \mathcal{L}$

$$E, s \models_i B_j \phi \Leftrightarrow \forall t \epsilon S\ ((sR_i t)\ implies\ E, t \models_j \phi)$$
$$E, s \dashv_i B_j \phi \Leftrightarrow \exists t \epsilon S\ ((sR_i t)\ and\ E, t \dashv_j \phi)$$

– $\forall E \epsilon \mathcal{E}, \forall s \epsilon S, \forall agents\, i, j,\ \forall \phi \epsilon \mathcal{L}$

$$E, s \models_i \neg B_j \phi \Leftrightarrow E, s \dashv_i B_j \phi$$
$$E, s \dashv_i \neg B_j \phi \Leftrightarrow E, s \models_i B_j \phi$$

A propositional formula ϕ is supported in a given situation s by an $Agent_i$ if and only if $Agent_i$ has reasons to think that ϕ is true in s. Analogously, ϕ will be rejected if and only if there are reasons that support its falsehood (recall that a formula may be both supported and rejected in a given situation). As far as beliefs are concerned, in a given situation s, $Agent_i$ supports that $Agent_j$ believes ϕ just in case $Agent_j$ supports ϕ in all the situations that are considered possible by $Agent_i$ in s ($Agent_i$'s doxastic alternatives). Similarly, $Agent_i$ may reject the fact that $Agent_j$ believes ϕ if it may think of a possible situation in which $Agent_j$ rejects ϕ. Finally, $Agent_i$ will support that $Agent_j$ does not believe ϕ if it may reject the fact that $Agent_j$ believes ϕ. We do not need more clauses to define the behaviour of the satisfiability and unsatisfiability relationships due to the restriction to linearly nested formulae imposed in definition 3.

4 Properties of the belief operators

The definition of an structure of subjective situations, the fact that the accessibility relations are equivalence relations and the clauses that describe the behaviour of the satisfiability (and unsatisfiability) relations compose a framework in which the modal belief operator of each $Agent_i$ has several interesting logical properties (that, in our opinion, make it an appropriate operator to model the notion of belief for a non-ideal agent). Some of these properties are described in this section.

4.1 General results

Proposition 1 (Lack of Logical Omniscience)

In the framework of subjective situations, none of the following forms of logical omniscience ([8]) holds:

- *Full logical omniscience.*
- *Belief of valid formulae.*
- *Closure under logical implication.*
- *Closure under logical equivalence.*
- *Closure under material implication.*
- *Closure under valid implication.*
- *Closure under conjunction.*
- *Weakening of beliefs.*
- *Triviality of inconsistent beliefs.*

Proof: Let us take a state s in which $T_i(s) = \{P, (P \to Q), \neg P\}$ and $F_i(s) = \{P\}$. Consider an structure for subjective situations E that only contains the situation s.

- $E, s \models_i B_i P$ and $E, s \models_i B_i(P \to Q)$ hold, but $E, s \models_i B_i Q$ does not hold. Therefore, neither full logical omniscience nor closure under material implication hold.
- $E, s \models_i B_i(Q \vee \neg Q)$ does not hold. Therefore, there is no belief of valid formulae.
- $E, s \models_i B_i P$ holds, but $E, s \models_i B_i(P \vee Q)$ does not hold. Therefore, closure under logical implication and weakening of beliefs do not hold.
- $E, s \models_i B_i(P \to Q)$ holds, but $E, s \models_i B_i(\neg Q \to \neg P)$ does not. Therefore, beliefs are not closed under logical equivalence or under valid implication.
- $E, s \models_i B_i P$ and $E, s \models_i B_i(P \to Q)$ hold, but the expression $E, s \models_i B_i(P \wedge (P \to Q))$ does not hold. Therefore, there is no closure under conjunction.
- $E, s \models_i B_i P$ and $E, s \models_i B_i \neg P$ hold, but $E, s \models_i B_i Q$ does not hold. Therefore, there is no triviality of inconsistent beliefs. $\square$

There are two basic reasons that account for the failure of all these properties:

- $\mathcal{T}_i$ and $\mathcal{F}_i$ are defined on sets of (arbitrary) formulae (not on basic propositions).
- $\mathcal{T}_i$ and $\mathcal{F}_i$ are unrelated. Thus, a given formula may belong to both sets, to only one of them or to none of them.

It is possible to impose any of the above properties on the belief operators by requiring these sets of formulae to satisfy some conditions (for instance, if $(\phi \wedge \psi) \epsilon \mathcal{T}_i(s)$ implies that $\phi \epsilon \mathcal{T}_i(s)$ and $\psi \epsilon \mathcal{T}_i(s)$, then $Agent_i$'s belief set would be closed under conjunction).

Proposition 2 (Relation between $\models_i$ and $\dashv_i$)

For any linearly nested formula ϕ
$E, s \not\models_i \phi$ *does not imply* $E, s \dashv_i \phi$
$E, s \models_i \phi$ *does not imply* $E, s \not\dashv_i \phi$

Proof: Take the structure of subjective situations E described in the proof of the previous proposition. It is easy to check these facts:

- $E, s \not\models_i B_i R$ and $E, s \not\models\!\!\!|_i B_i R$. Therefore, $E, s \not\models_i \phi$ does not imply $E, s \dashv_i \phi$.
- $E, s \models_i B_i P$ and $E, s \dashv_i B_i P$. Therefore, $E, s \models_i \phi$ does not imply $E, s \not\dashv_i \phi$. □

4.2 Results on positive introspection

Proposition 3 (Characterization of positive beliefs)
For any linearly nested formula ϕ,

$$E, s \models_i \phi \textit{ if and only if } E, s \models_i B_i\phi$$

Proof: The *if* side of the formula coincides with proposition 4. The *only if* side may be proven as follows:
$E, s \models_i B_i\phi \implies \forall t(sR_it), (E, t \models_i \phi)$. As R_i is reflexive, (sR_is); therefore, $E, s \models_i \phi$. □

This result states that $Agent_i$ believes ϕ in state s if and only if ϕ is one of the facts that is supported by $Agent_i$ in that state[5]. Thus, in our framework the difference between *belief* and *knowledge* vanishes: both concepts have to be understood as the propositional attitude that the agents adopt towards those formulae that they perceive to be true in the environment. Therefore, the (rather philosophical) difference between those beliefs that are true in the real world (that constitute knowledge) and those that are not (*plain* beliefs) is not taken into account.

Proposition 4 (Belief of supported formulae)

For any linearly nested formula ϕ,
$E, s \models_i \phi$ *implies* $E, s \models_i B_i\phi$

Proof: There are five cases to be considered:

- ϕ is a propositional formula.
  $E, s \models_i \phi$ and ϕ is propositional $\implies \phi \epsilon \mathcal{T}_i(s) \implies \forall t(sR_it), \phi \epsilon \mathcal{T}_i(t)$
  $\implies \forall t(sR_it), E, t \models_i \phi \implies E, s \models_i B_i\phi$
- If ϕ is a modal formula that starts with an affirmed belief operator B_i (*i.e.* $\phi = B_i\psi$), this fact is exactly the next proposition.
- If ϕ is a modal formula that starts with an affirmed belief operator B_j (*i.e.* $\phi = B_j\psi$), this statement coincides with proposition 6, that will be proved later.
- If ϕ is a modal formula that starts with a negated belief operator B_i (*i.e.* $\phi = \neg B_i\psi$), this fact is the one proved as proposition 10.
- If ϕ is a modal formula that starts with a negated belief operator B_j (*i.e.* $\phi = \neg B_j\psi$), this fact is the one proved as proposition 11. □

[5] The "only if" side of the proposition is the classical axiom of knowledge, axiom T.

This proposition is telling us that an agent believes all formulae that it has reasons to support, as suggested in the motivating example. However, this proposition has an added value over our intuitions, because it refers to any kind of linearly nested formulae, and not only to propositional formulae.

Proposition 5 (Single-agent positive introspection)

For any linearly nested formula ϕ,
$E, s \models_i B_i\phi$ *implies* $E, s \models_i B_iB_i\phi$

Proof: If $E, s \models_i B_i\phi$, that means that $E, s \models_i \phi$ holds in all the situations R_i-related to s. Being R_i an equivalence relation, these situations are exactly the ones included in the equivalence class of s induced by R_i. This class is also the set of situations that may be accessed from s in two steps (in fact, in any number of steps) *via* R_i, and ϕ is supported by $Agent_i$ in all of them. Thus, $\forall s'(sR_is')\forall s''(s'R_is'')E, s'' \models_i \phi$, and $E, s \models_i B_iB_i\phi$ also holds. □

This proposition states that axiom *4* (the classical axiom of positive introspection) holds for each belief operator B_i (*i.e.* every agent has introspective capabilities on its own positive beliefs).

Proposition 6 (Generation of positive beliefs)

$E, s \models_i B_j\phi$ *implies* $E, s \models_i B_iB_j\phi$

Proof: $E, s \models_i B_j\phi \implies \forall t(sR_it), E, t \models_j \phi$. Thus, $E, t \models_j \phi$ holds in all the worlds t that belong to the same equivalence class that s (considering the partition defined by R_i). Therefore, in all the worlds accessible from s *via* R_i in any number n of steps, $E, t \models_j \phi$. Taking the case $n = 2$, we obtain that $E, s \models_i B_iB_j\phi$. □

If an agent has reasons to support a certain belief of another agent, then that belief will be included in its belief set.

Proposition 7 (Inter-agent positive introspection)

$E, s \models_i B_j\phi$ *implies* $E, s \models_i B_jB_j\phi$

Proof: $E, s \models_i B_j\phi \implies \forall t(sR_it), E, t \models_j \phi$. Using the result given in proposition 4, that formula implies that $\forall t(sR_it), E, t \models_j B_j\phi$; thus, $E, s \models_i B_jB_j\phi$. □

This result is more general (proposition 5 reflected the case $i = j$). It states that each agent is aware of the fact that the other agents also have introspective capabilities.

Proposition 8 (Multi-agent positive introspection)

It does not hold (for three different agents $Agent_i$, $Agent_j$ and $Agent_k$ and a linearly nested formula ϕ) that

$E, s \models_i B_j\phi$ *implies* $E, s \models_i B_kB_j\phi$

Proof: We will show a counterexample. Take an structure for subjective situations E with two situations, s and t, such that (sR_kt) holds, but (sR_it) and (sR_jt) do not. Take a formula ϕ such that $\phi\epsilon\mathcal{T}_j(s)$ and $\phi\not\epsilon\mathcal{T}_j(t)$. In this state of affairs, $E, s \models_i B_j\phi$ holds but $E, s \models_i B_kB_j\phi$ does not hold. $\square$

This proposition states a negative result. It is telling that even if $Agent_i$ has reasons to support that $Agent_j$ believes something, that is not enough for $Agent_i$ to think that any other $Agent_k$ will have that belief. This proposition is essentially expressing the uncertainty of $Agent_i$ about the beliefs of a different $Agent_k$.

4.3 Results on negative introspection

Proposition 9 (Characterization of negative beliefs)
For any linearly nested formula ϕ,

$$E, s \mathrel{=\!\!|}_i \phi \textit{ if and only if } E, s \models_i \neg B_i\phi$$

Proof: The *if* side of the proposition may be proven as follows. As we know that $E, s \mathrel{=\!\!|}_i \phi$ and (sR_is), it may be said that $\exists t(sR_it), E, t \mathrel{=\!\!|}_i \phi$. Therefore, $E, s \mathrel{=\!\!|}_i B_i\phi$, which is equivalent to $E, s \models_i \neg B_i\phi$.

The *only if* side of the proposition (*i.e.* $E, s \models_i \neg B_i\phi$ implies $E, s \mathrel{=\!\!|}_i \phi$) will be proved considering five different cases (as we did in the proof of proposition 4):

- ϕ is a propositional formula.
  $E, s \models_i \neg B_i\phi \Longrightarrow E, s \mathrel{=\!\!|}_i B_i\phi \Longrightarrow \exists t(sR_it), E, t \mathrel{=\!\!|}_i \phi$. As ϕ is propositional, $E, t \mathrel{=\!\!|}_i \phi$ implies that $\phi\epsilon\mathcal{F}_i(t)$; as (sR_it), $\phi\epsilon\mathcal{F}_i(s)$. Therefore, $E, s \mathrel{=\!\!|}_i \phi$.
- ϕ is a modal formula that starts with an affirmed belief operator B_i (*i.e.* $\phi = B_i\psi$).

  $$E, s \models_i \neg B_i\phi \Longrightarrow E, s \models_i \neg B_iB_i\psi \Longrightarrow E, s \mathrel{=\!\!|}_i B_iB_i\psi \Longrightarrow$$
  $$\exists t(sR_it), E, t \mathrel{=\!\!|}_i B_i\psi \Longrightarrow \exists t, u(sR_it), (tR_iu), E, u \mathrel{=\!\!|}_i \psi.$$

  As R_i is transitive, (sR_it) and (tR_iu) imply that (sR_iu). Thus, we may state that $\exists u(sR_iu), E, u \mathrel{=\!\!|}_i \psi$. Therefore, $E, s \mathrel{=\!\!|}_i B_i\psi$, which is equal to $E, s \mathrel{=\!\!|}_i \phi$.
- ϕ is a modal formula that starts with an affirmed belief operator B_j (*i.e.* $\phi = B_j\psi$).

  $$E, s \models_i \neg B_i\phi \Longrightarrow E, s \models_i \neg B_iB_j\psi \Longrightarrow E, s \mathrel{=\!\!|}_i B_iB_j\psi \Longrightarrow$$
  $$\exists t(sR_it), E, t \mathrel{=\!\!|}_i B_j\psi \Longrightarrow \exists t, u(sR_it), (tR_iu), E, u \mathrel{=\!\!|}_j \psi.$$

  As R_i is transitive, (sR_it) and (tR_iu) imply that (sR_iu). Thus, we may state that $\exists u(sR_iu), E, u \mathrel{=\!\!|}_j \psi$. Therefore, $E, s \mathrel{=\!\!|}_i B_j\psi$, which is equal to $E, s \mathrel{=\!\!|}_i \phi$.
- ϕ is a modal formula that starts with a negated belief operator B_i (*i.e.* $\phi = \neg B_i\psi$).

  $$E, s \models_i \neg B_i\phi \Longrightarrow E, s \models_i \neg B_i\neg B_i\psi \Longrightarrow E, s \mathrel{=\!\!|}_i B_i\neg B_i\psi \Longrightarrow$$
  $$\exists t(sR_it), E, t \mathrel{=\!\!|}_i \neg B_i\psi \Longrightarrow \exists t(sR_it), E, t \models_i B_i\psi \Longrightarrow$$
  $$\exists t(sR_it)\forall u(tR_iu), E, u \models_i \psi.$$

In this expression, t is a world that belongs to the same class of equivalence than s (according to the partition defined by R_i), and u represents all the worlds that belong to t's class of equivalence; thus, u ranges over all the worlds belonging to s's class of equivalence (all the worlds that are accessible from s via R_i in any number n of steps). If we take $n = 1$, we get that $\forall t(sR_it), E, t \models_i \psi$. Thus, $E, s \models_i B_i\psi$, which is equivalent to $E, s \rightmodels_i \neg B_i\psi$. Therefore, $E, s \rightmodels_i \phi$.

– ϕ is a modal formula that starts with a negated belief operator B_j (*i.e.* $\phi = \neg B_j\psi$).

$$E, s \models_i \neg B_i\phi \Longrightarrow E, s \models_i \neg B_i\neg B_j\psi \Longrightarrow$$
$$E, s \rightmodels_i B_i\neg B_j\psi \Longrightarrow \exists t(sR_it), E, t \rightmodels_i \neg B_j\psi \Longrightarrow$$
$$\exists t(sR_it), E, t \models_i B_j\psi \Longrightarrow \exists t(sR_it)\forall u(tR_iu), E, u \models_j \psi.$$

In this expression, t is a world that belongs to the same class of equivalence than s (according to the partition defined by R_i), and u represents all the worlds that belong to t's class of equivalence; thus, u ranges over all the worlds belonging to s's class of equivalence (all the worlds that are accessible from s via R_i in any number n of steps). If we take $n = 1$, we get that $\forall t(sR_it), E, t \models_j \psi$. Thus, $E, s \models_i B_j\psi$, which is equivalent to $E, s \rightmodels_i \neg B_j\psi$. Therefore, $E, s \rightmodels_i \phi$. □

$Agent_i$ does not believe ϕ at s if and only if ϕ is one the facts that is rejected by i at s. Again, this proposition agrees with the intuitions that we had in the example that was used to motivate the need for the framework of subjective situations.

Proposition 10 (Single-agent negative introspection)

$$E, s \models_i \neg B_i\phi \textit{ implies } E, s \models_i B_i\neg B_i\phi$$

Proof: $E, s \models_i \neg B_i\phi \Longrightarrow E, s \rightmodels_i B_i\phi \Longrightarrow \exists t(sR_it), (E, t \rightmodels_i \phi)$. Thus, there exists at least one world (say w) such that (sR_iw) and $E, w \rightmodels_i \phi$. In order to prove the proposition, we have to notice that R_i is Euclidean (*i.e.* whenever (sR_it) and (sR_iu), (tR_iu) also holds)[6]. Therefore, w is R_i accessible from all worlds that are R_i accessible from s, and we may state that $\forall t(sR_it), (tR_iw)$ and $E, w \rightmodels_i \phi$. Thus, $\forall t(sR_it)\ \exists u(tR_iu)E, u \rightmodels_i \phi$. Thus, $\forall t(sR_it)\ E, t \rightmodels_i B_i\phi$, which is equivalent to $\forall t(sR_it)\ E, t \models_i \neg B_i\phi$. Therefore, we have shown that $E, s \models_i B_i\neg B_i\phi$. □

This proposition states that axiom *5* (the classical axiom of negative introspection) holds for each belief operator B_i (*i.e.* every agent has introspective capabilities on its own negative beliefs).

Proposition 11 (Generation of negative beliefs)

$$E, s \models_i \neg B_j\phi \textit{ implies } E, s \models_i B_i\neg B_j\phi$$

Proof: $E, s \models_i \neg B_j\phi \Longrightarrow E, s \rightmodels_i B_j\phi \Longrightarrow \exists t(sR_it), E, t \rightmodels_j \phi$. Let us call w to any of the worlds referred to by this existential quantifier. Being R_i Euclidean, we know that $\forall t(sR_it), (tR_iw)$; therefore, we may say that $\forall t(sR_it)\exists u(tR_iu), E, u \rightmodels_j \phi$. Thus, $\forall t(sR_it), E, t \rightmodels_i B_j\phi$, which is equivalent to $\forall t(sR_it), E, t \models_i \neg B_j\phi$. Therefore, $E, s \models_i B_i\neg B_j\phi$. □

This proposition is expressing the fact that $Agent_i$ can make positive introspection on negated beliefs of other agents.

[6] It is easy to prove that any relation that is symmetric and transitive is also Euclidean.

4.4 Summary of the main properties

Summarising the main results shown in this section:

- All forms of logical omniscience are avoided.
  None of the restricted forms of logical omniscience usually considered in the literature holds in the framework of subjective situations. This result is due to the presence of partial and inconsistent situations and to the fact that the description of a situation is formed with positive and negative information about propositional formulae (and not about basic propositions).
- Each agent is aware of its positive and negative beliefs, and is also aware of the fact that the other agents enjoy this introspective capability.
  However, an agent is uncertain about the way the present situation is perceived by other agents and, therefore, it is unable to know anything about the other agent's beliefs.
- The positive and negative beliefs of an agent in an state reflect, as our intuitions suggested, the facts that are taken as true or false by the agent in that state.
  Thus, an agent's perception determines its beliefs in a given situation, as it might be expected.

5 Comparison with previous proposals

The most outstanding difference of our proposal with previous works ([8]) is the idea of considering *subjective* situations, that may be perceived in different ways by different agents. Technically, this fact implies two differences of our approach with respect to others:

- A situation is described with two functions ($\mathcal{T}_i$ and $\mathcal{F}_i$) *for each* $Agent_i$.
  Thus, we take into account each agent's perception of the actual situation, considering a *subjective* description of each state.
- Two satisfiability and unsatisfiability relations between situations and formulae ($\models_i$ and $=\!\!|_i$) are also defined *for each agent*.
  Having a *subjective* description of each state, it makes sense to consider satisfiability relations that depend on each agent.

The rest of the section is devoted to the comparison of our proposal with the two approaches to the problem of logical omniscience with which it shares more similarities: Levesque's *logic of explicit and implicit beliefs* ([6]) and Thijsse's *hybrid sieve systems* ([7]).

5.1 Levesque's logic of implicit and explicit beliefs

Levesque uses a language with two modal operators: *B* for *explicit* beliefs and *L* for *implicit* beliefs. These operators are not allowed to be nested in the formulae of the language. A *structure for explicit and implicit beliefs* is defined as a tuple *M*=(*S*, $\mathcal{B}$, *T*, *F*), where *S* is the set of primitive situations, $\mathcal{B}$ is a subset of *S* that represents the situations that could be the actual one and T and F are functions from the set of primitive propositions into subsets of S. Intuitively, $T(P)$ contains all the situations that support the truth of P, whereas $F(P)$ contains the ones that support the falsehood of P. A situation s can be *partial* (if there is a primitive proposition which is neither true nor false in s) and/or *incoherent* (if there is a proposition which is both true and false in s). A situation is *complete* if it is neither partial nor incoherent. A complete situation s is *compatible* with a situation t if s and t agree in all the points in which t is defined. $\mathcal{B}^*$ is the set of all complete situations of S that are compatible with some situation in $\mathcal{B}$.

The relations $\models_T$ and $\models_F$ between situations and formulae are defined as follows:

- $M,s \models_T P$, where P is a primitive proposition, if and only if $s \in T(P)$
- $M,s \models_F P$, where P is a primitive proposition, if and only if $s \in F(P)$
- $M,s \models_T \neg\varphi$ if and only if $M,s \models_F \varphi$
- $M,s \models_F \neg\varphi$ if and only if $M,s \models_T \varphi$
- $M,s \models_T (\varphi \wedge \psi)$ if and only if $M,s \models_T \varphi$ and $M,s \models_T \psi$
- $M,s \models_F (\varphi \wedge \psi)$ if and only if $M,s \models_F \varphi$ or $M,s \models_F \psi$
- $M,s \models_T B\varphi$ if and only if $M,t \models_T \varphi \;\; \forall t \epsilon \mathcal{B}$
- $M,s \models_F B\varphi$ if and only if $M,s \not\models_T B\varphi$
- $M,s \models_T L\varphi$ if and only if $M,t \models_T \varphi \;\; \forall t \epsilon \mathcal{B}^*$
- $M,s \models_F L\varphi$ if and only if $M,s \not\models_T L\varphi$

There are some similarities between our approach and Levesque's logic of implicit and explicit beliefs. However, they are more apparent than real, as shown in this listing:

- Levesque also considers a satisfiability and an unsatisfiability relation between situations and doxastic formulae.
  However, these relations are not considered *for each agent*.
- Levesque also describes each situation with two functions $\mathcal{T}$ and $\mathcal{F}$.
  These functions are not indexed by each agent, as our functions are (Levesque considers an objective description of what is true and what is false in each situation). Another important difference is that Levesque's functions deal with basic propositions, and not with formulae as our functions do.
- Both approaches allow the presence of *partial* or *inconsistent* situations.
  However note that, in our case, it is not the (objective) description of the situation that is partial or inconsistent, but the *subjective* perception that an agent may have of it. Thus, the notions of partiality and inconsistency have a much more natural interpretation in our framework.
- Both approaches avoid all the forms of logical omniscience.
  The reason is different in each case, though. In Levesque's logic of explicit and implicit beliefs, it is the presence of incoherent situations that prevents logical omniscience. In our proposal, there is no need to have inconsistent situations to avoid logical omniscience. In fact, we solve that problem by defining $\mathcal{T}_i$ and $\mathcal{F}_i$ over arbitrary sets of formulae, and not over basic propositions.
- There are accessibility relations between situations for each agent in both systems.
  Levesque's accessibility relation between situations is left implicit; our accessibility relations are explicit. Furthermore, the intuition underlying these relations is somewhat different, as explained in section 2.

Other differences with Levesque's approach are:

- Levesque only considers one agent, and does not allow nested beliefs. Thus, his agents do not have any introspective capabilities.
- Levesque defines *explicit* and *implicit* beliefs, whereas we do not make this distinction.
- Even though Levesque avoids logical omniscience, his agents must necessarily believe all those tautologies that are formed by *known* basic propositions (those propositions P for which the agent believes $(P \vee \neg P)$), regardless of their complexity. This is not the case in our approach, because we deal directly with formulae.
- There is a different treatment of the unsatisfiability relation when applied to beliefs, because he transforms $=\!\mid$ into $\not\models$, whereas we do not.

5.2 Thijsse's hybrid sieve systems

Thijsse ([7]) proposes a way of using *partial logics* to deal with various forms of logical omniscience. He defines a *partial model* as a tuple $(W, \mathcal{B}_1, \ldots, \mathcal{B}_n, V)$, where W is a set of worlds, $\mathcal{B}_i$ is the accessibility relation between worlds for $Agent_i$ and V is a *partial* truth assignment to the basic propositions in each world. $\top$ is a primitive proposition that is always interpreted as *true*. Truth ($\models$) and falsity ($=\!\mid$) relations are defined in the following way:

- $M,w \models \top$
- $M,w \not=\!\mid \top$
- $M,w \models P$, where P is a primitive proposition, iff $V(P,w)$= 1
- $M,w =\!\mid P$, where P is a primitive proposition, iff $V(P,w)$= 0
- $M,w \models \neg\varphi$ iff $M,w =\!\mid \varphi$
- $M,w =\!\mid \neg\varphi$ iff $M,w \models \varphi$
- $M,w \models (\varphi \wedge \psi)$ iff $M,w \models \varphi$ and $M,w \models \psi$
- $M,w =\!\mid (\varphi \wedge \psi)$ iff $M,w =\!\mid \varphi$ or $M,w =\!\mid \psi$
- $M,w \models B_i\varphi$ iff $M,v \models \varphi\ \forall v$ such that $(w, v)\ \epsilon \mathcal{B}_i$
- $M,w =\!\mid B_i\varphi$ iff $\exists v$ s.t. $(w, v)\ \epsilon \mathcal{B}_i$ and $M,v =\!\mid \varphi$

The most important similarities between our approach and Thijsse's are:

- n agents and n explicit accessibility relations are considered.
  However, as in Levesque's case, there are no restrictions on these relations, and the intuitive meaning of our accessibility relations is slightly different.
- Two relations (of satisfiability and unsatisfiability) are defined. Moreover, a similar clause is used to provide a meaning to the unsatisfiability relation with respect to the belief operator.
  As before, the main difference is that we provide two relations *for each agent*.
- There are no tautologies in Thijsse's system; therefore, he does not have to care about some forms of logical omniscience (closure under valid implication and belief of valid formulae).
- Closure under material implication and closure under conjunction do not hold in Thijsse's approach either.

The main difference with Thijsse's proposal is that he uses *partial* assignments of truth values *over basic propositions* for each state; thus, a proposition may be true, false or undefined in each state. We deal with formulae, not with basic propositions, and each formula may be supported *and/or* rejected by *each agent* in each state. Therefore, Thijsse's approach is three-valued, whereas ours is more of a four-valued kind, such as Levesque's.

6 Summary

In this paper it has been argued that each agent perceives its actual situation in a particular way, which may be different from that of other agents located in the same situation. The vision that an agent has of a situation determines its (positive and negative) beliefs in that situation. This intuitive idea has been formalized with the notion of *subjective situations*. These entities are the base of a doxastic logic, in which the meaning of the belief operators seems to fit with the general intuitions about how the doxastic attitude of a non-ideal agent should behave. In particular, logical omniscience is avoided while some interesting introspective properties are maintained. A detailed comparison of this approach with Levesque's *logic of implicit and explicit beliefs* ([6]) and Thijsse's *hybrid sieve systems* ([7]) has also been provided.

References

1. W. van der Hoek, J.-J. Meyer, *Epistemic logic for AI and Computer Science*, Cambridge Tracts in Theoretical Computer Science 41, Cambridge University Press, 1995.
2. J. Hintikka, *Knowledge and belief*, Cornell University Press, Ithaca, N.Y., 1962.
3. S. Kripke, A semantical analysis of modal logic I: normal modal propositional calculi, *Zeitschrift für Mathematische Logik und Grundlagen Mathematik* **9** (1963), 67-96.
4. J. Hintikka, Impossible possible worlds vindicated, *J. of Phil. Logic* **4** (1975), 475-484.
5. A.Moreno, U.Cortés, T.Sales, *Subjective situations*, in Multi-Agent System Engineering, F.J.Garijo and M.Boman (Eds.), Lecture Notes in Artificial Intelligence 1647, pp. 210-220, 1999.
6. H.J. Levesque, A logic of implicit and explicit belief, *Proceedings of the Conference of the American Association for Artificial Intelligence, AAAI-84* (1984), 198-202.
7. E. Thijsse, Combining partial and classical semantics. A hybrid approach to belief and awareness, in P. Doherty (Ed.), *Partiality, modality, and nonmonotonicity*, Studies in Logic, Language and Information, Center for the Study of Language and Information Publications (1996), 223-249.
8. A. Moreno, Avoiding logical omniscience and perfect reasoning: a survey, *AI Communications* **2** (1998), 101-122.
9. A. Moreno, T. Sales, Dynamic belief analysis, in J. Müller, M. Wooldridge, N. Jennings (Eds.), *Intelligent Agents III: Agent theories, architectures and languages*, Lecture Notes in Artificial Intelligence 1193, Springer Verlag (1997), 87-102.
10. A.Moreno, U.Cortés, T.Sales, Inquirers: a general model of non-ideal agents, *International Journal of Intelligent Systems* 15 (3), pp. 197-215, 2000.

This article was processed using the LaTeX macro package with LLNCS style