Guillaume Aucher

An internal version of epistemic logic

Abstract. Representing an epistemic situation involving several agents obviously depends on the modeling point of view one takes. We start by identifying the types of modeling points of view which are logically possible. We call the one traditionally followed by epistemic logic the perfect external approach, because there the modeler is assumed to be an omniscient and external observer of the epistemic situation. In the rest of the paper we focus on what we call the internal approach, where the modeler is one of the agents involved in the situation. For this approach we propose and axiomatize a logical formalism based on epistemic logic. This leads us to formalize some intuitions about the internal approach and about its connections with the external ones. Finally, we show that our internal logic is decidable and PSPACE-complete.

Keywords: Epistemic logic, multi-agent systems, internal approach.

1. Introduction

In the literature about epistemic logic, when it comes to model epistemic situations, not much is said explicitly about which modeling point of view is considered. However, modeling an epistemic situation depends very much on the modeling point of view. Indeed, the models built to represent the situation will be quite different depending on whether the modeler is an agent involved in the situation or not. To illustrate this point, let us consider the following example. Assume that the agents Yann and Alice are in a room and that there is a coin in a box that both cannot see because the box is closed. Now, assume that Alice cheats, opens the box and looks at the coin. Moreover, assume that Yann does not suspect anything about it and that Alice knows it (Yann might be inattentive or out of the room for a while). How can we represent this resulting situation? On the one hand, if the modeler is an external observer (different from Yann and Alice) knowing everything that has happened, then in the model that this external observer builds Alice knows whether the coin is heads or tails up. On the other hand, if the modeler is Yann then in the model that Yann builds Alice does *not* know whether the coin is heads or tails up. As we see in this example, the intuitive interpretation of a model really makes sense only when one knows the modeling point of view.

The importance of specifying a modeling point of view is also stressed at a great extent in Newtonian mechanics in physics where physicists must al-

ways specify which *frame of reference* they consider when they want to study a natural phenomenon. And just as for epistemic situations, the representation of this phenomenon depends very much on this frame of reference. For example, assume somebody drops a ball from the top of the high mast of a ship sailing nearby a harbor. Then, viewed from the frame of reference of the ship, the trajectory of the ball will be a straight line. But viewed from the frame of reference of the harbor, the trajectory will be a parabola (the more rapidly the ship sails and the higher the mast is, the more curved the parabola will be).

Given an epistemic situation, assume that we want to model the beliefs of the agents $G = \{j_1, \ldots, j_N\}$ and possibly the actual state of the world. What kinds of modeling points of view are there? For a start, we can distinguish whether or not the modeler is one of these agents G under scrutiny.

1. First, consider the case where the modeler is *one* of the agents G. In the rest of the paper we call this modeler-agent agent Y (like You). The models she builds could be seen as models she has 'in her mind'. They represent the way she perceives the surrounding world. In that case, agent Y is involved in the situation, she is considered on a par by the other agents and interacts with them. So she should be represented in the formalism and her models should deal not only with the other agents' beliefs but also with the other agents' beliefs about her own beliefs. This is an internal and subjective point of view, the situation is modeled from the inside. Therefore, for this very reason her beliefs might be erroneous. Hence the models she builds might also be erroneous. We call this agent point of view the *internal* point of view.
2. Second, consider the case where the modeler is *not* one of the agents G. The modeler is thus an observer external to the situation. She is not involved in the situation and she does not exist for the agents, or at least she is not taken into consideration in their representation of the world. So she should not be represented in the formalism and particularly the agents' beliefs about her own beliefs should also not be represented. The models that this modeler builds are supposed to represent the situation 'from above', from an external and objective point of view. There are then two other sub-possibilities depending on whether or not the modeler has a perfect knowledge of the situation.
   (a) In case the modeler has a perfect knowledge of the situation, then everything that is true in the model that she builds is true in reality

	the modeler is uncertain about the situation	the modeler is one of the agents
internal approach	•	•
imperfect external approach	•	
perfect external approach		

Figure 1. Essential differences between the internal and external approaches

and vice versa, everything that is true in reality is also true in the model. This thesis was already introduced in [5]. Basically, the models built by the modeler are perfectly correct. The modeler has access to the minds of the agents and knows perfectly not only what they believe but also what the actual state of the world is. This is a kind of 'divine' point of view and we call it the *perfect external* point of view.

(b) In case the modeler does not have a perfect knowledge of the situation then, like the internal point of view but unlike the perfect external point of view, the models built might be erroneous. The models could also be correct but then, typically, the modeler would be uncertain about which is the actual world (in that sense, she would not have a perfect knowledge of the situation). What the modeler knows can be obtained for example by observing what the agents say and do, by asking them questions... We call this point of view the *imperfect external* point of view.

Because we proceeded by successive dichotomies, we claim that the internal, the perfect external and the imperfect external points of view are the only three logically possible points of view when we want to model epistemic situations. From now on we will call these modeling approaches the internal, the external and the imperfect external approaches; their differences are summarized in Figure 1.* The fields of application of these three approaches are different. The internal and imperfect external approaches have rather applications in artificial intelligence where agents/robots acting in the world

*In [24], the internal and external points of view are studied from a broader philosophical perspective and not just for their need in representing agents' beliefs. Nagel mainly deals there with the issues of how these views can be combined and if they can possibly be integrated. He does so by tracing the manifestations of these issues in a number of philosophical topics: the metaphysics of mind, the theory of knowledge, free will, and ethics. He argues that giving a complete account of reality (as in philosophy of mind) or of all reasons for actions (as in ethics) in objective terms *only* is not always possible.

need to have a formal representation of the surrounding world and to cope with uncertain information. The internal approach has also applications in cognitive psychology where the aim is to model the cognition of one agent (possibly in a multi-agent setting). The perfect external approach has rather applications in game theory [7], social psychology (distributed cognition) or distributed systems [14] for example. Indeed, in these fields we need to model situations accurately from an external point of view in order to explain and predict what happens in these situations.

The modeling point of view is definitely not the only important factor to specify when one wants to model epistemic situations: the second important factor is obviously our *object of study*, i.e. *what* we actually model. Typically, it is the actual state of the world and the beliefs of the agents G about each other. But this could also perfectly be their beliefs about other agents $j'_1, \ldots, j'_{N'}$ or the beliefs of only *some* of these agents G (about *all* the agents G) for instance. Therefore, to proceed methodically and properly (and similarly as in physics), when one wants to model epistemic situations one should ideally specify from the start a combination of these two factors. Indeed, each combination gives rise to a particular kind of formalism. However some combinations might turn out to be equivalent to others: for example, if the object of study is the epistemic state of a single agent Y (in a single or a multi-agent setting), then the perfect external approach for this object of study amounts to the internal approach where the modeler-agent is Y herself and the object of study is the actual state of the world (and possibly the other agents' beliefs about each other in a multi-agent setting). This example suggests that the internal approach is somehow reducible to the perfect external approach if we specify appropriate objects of study. But because the corresponding object of study in the external approach of a given object of study in the internal approach might be quite convoluted in some cases we prefer to keep the natural and intuitive distinction between the internal and the perfect external approaches.

In the logical literature, some combinations have been already studied. In [6, 25, 9], the authors follow the imperfect external approach in a single agent setting and model only the epistemic state of this single agent. Standard epistemic logic [18] follows the external approach in a multi-agent setting and models the epistemic states of *all* the agents together with the actual state of the world. On the other hand, AGM belief revision theory [1] follows the internal approach but in a single agent setting. In fact there is no logical formalism that follows the internal approach in a multi-agent setting where the modeler-agent Y models the epistemic states of all the agents

together with the actual state of the world. However, such a formalism is crucial if we want to design autonomous agents for instance. That is what we are going to propose in this paper.

The paper is organized as follows. In Section 2 we recall epistemic logic. In Section 3 we propose a semantics for the internal approach in a multi-agent setting. Then in Section 4 we set some connections between the internal and the external approaches. Finally, in Section 5 we propose an axiomatization of the internal semantics.

NOTE 1.1. All the proofs of this paper can be found at the address `http://StableAddress`

2. Epistemic logic

Epistemic logic is a modal logic [8] that is concerned with the logical study of the notions of knowledge and belief. So what we call an epistemic model is just a particular kind of Kripke model as used in modal logic. The only difference is that instead of having a single accessibility relation we have a set of accessibility relations, one for each agent. This set of agents is noted G and its cardinality N. Besides, Φ is a set of propositional letters.

DEFINITION 2.1. An *epistemic model* M is a triple $M = (W, R, V)$ such that

- W is a non-empty set of possible worlds;
- $R : G \to 2^{W \times W}$ assigns an accessibility relation to each agent;
- $V : \Phi \to 2^W$ assigns a set of possible world to each propositional letter.

If $M = (W, R, V)$ is an epistemic model, a pair (M, w_a) with $w_a \in W$ is called a *pointed epistemic model*. We also note $R_j = R(j)$ and $R_j(w) = \{w' \in W; wR_jw'\}$, and $w \in M$ for $w \in W$.

Intuitively, a pointed epistemic model (M, w_a) represents from an external point of view how the actual world w_a is perceived by all the agents G. This entails that epistemic logic clearly follows the perfect external approach. The possible worlds W are the relevant worlds needed to define such a representation and the valuation V specifies which propositional facts (such as 'it is raining') are true in these worlds. Finally the accessibility relations R_j model the notion of belief. We set $w' \in R_j(w)$ in case in world w, agent j considers the world w' possible.

Finally, the submodel of M generated by a set of worlds $S \subseteq M$ is the restriction[†] of M to the worlds $\{(\bigcup_{j\in G} R_j)^*(w_S); w_S \in S\}$ (where $(\bigcup_{j\in G} R_j)^*$ is the reflexive transitive closure of $(\bigcup_{j\in G} R_j)$, see [8] for details). In case the submodel of M generated by a set of worlds $S \subseteq M$ is M itself then M is said to be generated by S and in case S is a singleton it is called the root of M. Intuitively, the submodel of M generated by a set of worlds S contains all the relevant information in M about these worlds S.

Now inspiring ourselves from modal logic, we can define a language for epistemic models which will enable us to express things about them. The modal operator is then a 'belief' operator, one for each agent.

Definition 2.2. The language $\mathcal{L}$ is defined as follows:

$$\mathcal{L} : \phi ::= \top \mid p \mid \neg\phi \mid \phi \wedge \phi \mid B_j\phi$$

where p ranges over Φ and j over G. Moreover, $\phi \vee \phi'$ is an abbreviation for $\neg(\neg\phi \wedge \neg\phi')$; $\phi \rightarrow \phi'$ is an abbreviation for $\neg\phi \vee \phi'$; $\hat{B}_j\phi$ is an abbreviation for $\neg B_j \neg\phi$; and $\bot$ is an abbreviation for $\neg\top$.

Now we can give meaning to the formulas of this language by defining truth conditions for these formulas on the class of epistemic models.

Definition 2.3. Let $M = (W, R, V)$ be an epistemic model and $w \in W$. $M, w \models \phi$ is defined inductively as follows:

$$\begin{array}{lll} M, w \models \top & & \\ M, w \models p & \text{iff} & w \in V(p) \\ M, w \models \neg\phi & \text{iff} & \text{not } M, w \models \phi \\ M, w \models \phi \wedge \phi' & \text{iff} & M, w \models \phi \text{ and } M, w \models \phi' \\ M, w \models B_j\phi & \text{iff} & \text{for all } v \in R_j(w), M, v \models \phi \end{array}$$

When $M, w \models \phi$, we say that ϕ *true* in w or ϕ is *satisfied* in w. We write $M \models \phi$ when $M, w \models \phi$ for all $w \in M$.

So agent j believes ϕ in world w (formally $M, w \models B_j\phi$) if ϕ is true in all the worlds that the agent j considers possible (in world w).

[†]Let $M = (W, R, V)$ be an epistemic model. The restriction of M to a set of worlds S is the submodel $M' = (W', R', V')$ of M defined as follows. $W' = W \cap S$; $R'_j = R_j \cap (S \times S)$ for all $j \in G$; and $V'(p) = V(p) \cap S$ for all $p \in \Phi$.

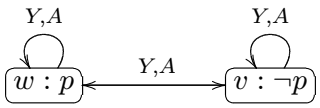

Figure 2. Epistemic model

EXAMPLE 2.4. In the pointed epistemic model (M, w) of Figure 2, p stands for 'the coin is heads up'. Y stands for Yann and A for Alice. Accessibility relations are represented by arrows. So Yann does not know whether the coin is heads or tails up: $M, w \models \neg B_Y p \wedge \neg B_Y \neg p$. Yann also believes that Alice does not know neither: $M, w \models B_Y(\neg B_A p \wedge \neg B_A \neg p)$. Finally, Yann believes that Alice believes that he does not know whether the coin is heads or tails up: $M, w \models B_Y B_A(\neg B_Y \neg p \wedge \neg B_Y p)$.

But note that the notion of belief might comply to some constraints (or axioms) such as $B_j\phi \rightarrow B_j B_j \phi$: if agent j believes something, she 'knows' that she believes it. These constraints might affect the nature of the accessibility relations R_j which may then comply to some extra properties. We list here some properties that will be useful in the sequel: seriality: for all w, $R_j(w) \neq \emptyset$; transitivity: for all w, w', w'', if $w' \in R_j(w)$ and $w'' \in R_j(w')$ then $w'' \in R_j(w)$; euclidicity: for all w, w', w'', if $w' \in R_j(w)$ and $w'' \in R_j(w)$ then $w' \in R_j(w'')$. Then we define the class of $\mathsf{KD45_G}$-models as the class of epistemic models whose accessibility relations are serial, transitive and euclidean. If for all $\mathsf{KD45_G}$-models M, $M \models \phi$ then ϕ is said to be $\mathsf{KD45_G}$-valid and it is noted $\models_{KD45_G} \phi$.

DEFINITION 2.5. The logic $\mathsf{KD45_G}$ is defined by the following axiom schemes and inference rules:

Taut	$\vdash_{KD45_G} \phi$ for all propositional tautologies ϕ
K	$\vdash_{KD45_G} B_j(\phi \rightarrow \psi) \rightarrow (B_j\phi \rightarrow B_j\psi)$ for all $j \in G$
D	$\vdash_{KD45_G} B_j\phi \rightarrow \hat{B}_j\phi$
4	$\vdash_{KD45_G} B_j\phi \rightarrow B_j B_j \phi$
5	$\vdash_{KD45_G} \neg B_j\phi \rightarrow B_j \neg B_j \phi$
Nec	If $\vdash_{KD45_G} \phi$ then $\vdash_{KD45_G} B_j\phi$ for all $j \in G$
MP	If $\vdash_{KD45_G} \phi$ and $\vdash_{KD45_G} \phi \rightarrow \psi$ then $\vdash_{KD45_G} \psi$.

We write $\vdash_{KD45_G} \phi$ in case ϕ belongs to the logic $\mathsf{KD45_G}$ and we say that ϕ is *provable* in $\mathsf{KD45_G}$.

An interesting feature of epistemic (and modal) logic is that we can somehow match the constraints imposed by the axioms on the belief operator B_j with constraints on the accessibility relations R_j. In other words, the notions of validity and provability coincide. That is what the following theorem expresses.

THEOREM 2.6 (soundness and completeness). *For all* $\phi \in \mathcal{L}$*,*

$$\vdash_{KD45_G} \phi \textit{ iff } \models_{KD45_G} \phi$$

As we said, epistemic logic rather follows the perfect external approach. So now we are going to propose a formalism for the internal approach.

3. A semantics for the internal approach

To define a semantics for the internal approach in a multi-agent setting, we will start from the standard view of an agent's epistemic state as a set of possible worlds (used in the AGM framework), and then extend it to the multi-agent case.

But first we have to make some assumption. As we said in the previous section, the internal approach has applications in artificial intelligence and in cognitive psychology. So the objects we introduce should be essentially finite. Indeed, computers cannot easily deal with infinite structures and a human cognition is by nature finite. So the set Φ of propositional letters is assumed to be finite.

3.1. Multi-agent possible world and internal model

In the AGM framework, one considers a single agent Y. The possible worlds are supposed to represent how the agent Y perceives the surrounding world. As she is the only agent, these possible worlds deal only with propositional facts about the surrounding world. Now, if we suppose that there are other agents than agent Y, a possible world for Y in that case should also deal with how the other agents perceive the surrounding world. These "multi-agent" possible worlds should then not only deal with propositional facts but also with epistemic facts. So to represent a multi-agent possible world we need to add a modal structure to our (single agent) possible worlds. We do so as follows.

DEFINITION 3.1. A *multi-agent possible world* (M, w) is a *finite* pointed epistemic model $M = (W, R, V, w)$ generated by $w \in W$ such that R_j is serial, transitive and euclidean for all $j \in G$, and

1. $R_Y(w) = \{w\}$;
2. there is no v and $j \neq Y$ such that $w \in R_j(v)$.

Let us have a closer look at the definition. Condition 2 will be motivated later (after Definition 3.6), but note that any pointed epistemic model satisfying the conditions of a multi-agent possible world except condition 2 is bisimilar to a multi-agent possible world. Condition 1 ensures that in case Y is the only agent then a multi-agent possible world boils down to a possible world, as in the AGM theory. Condition 1 also ensures that in case Y assumes that the situation is correctly represented by the multi-agent possible world (M, w) then for her w is the (only) actual world. In fact the other possible worlds of a multi-agent possible world are just present for technical reasons: they express the other agents' beliefs (in world w). One could get rid of the condition that a multi-agent possible world (M, w) is generated by w but the worlds which do not belong to the submodel generated by w would not have neither philosophical nor technical motivation. Besides, for the same reason that Φ is finite, a multi-agent possible world is also assumed to be finite. Finally, notice that we assume that accessibility relations are serial, transitive and euclidean. This means that the agents' beliefs are consistent and that agents 'know' what they believe and disbelieve (axioms D, 4 and 5 of Definition 2.5). These seem to be very natural constraints to impose on the notion of belief. Intuitively, this notion of belief corresponds for example to the kind of belief in a theorem that you have after having proved this theorem and checked the proof several times. In the literature, this notion of belief corresponds to Lenzen's notion of conviction [19] or to Gardenfors' notion of acceptance [15] or to Voorbraak's notion of rational introspective belief [28]. In fact, in all the agent theories the notion of belief satisfies these constraints: in Cohen and Levesque's theory of intention [11] or in Rao and Georgeff BDI architecture [16] [26] or in Meyer et. al. KARO architecture [27] [23] or in Wooldridge BDI logic LORA [29]. However, one should note that all these agent theories follow the perfect external approach. This is of course at odds with their intention to implement their theories in machines.

Remark 3.2. In this paper we deal only with the notion of belief but one could add the notion of knowledge and also easily the notion of common belief. Indeed, it might be interesting to express things such as 'agent Y believes that agent j does not *know* p' (even if this could be rephrased in terms of beliefs). However, note that in an internal approach agent Y's (proper) beliefs coincide with her knowledge. We refrain to introduce these notions in order to highlight the main new ideas and because in most applications of the internal approach the notion of knowledge is not essential. Nevertheless

a (single-agent) possible world:

a multi-agent possible world:

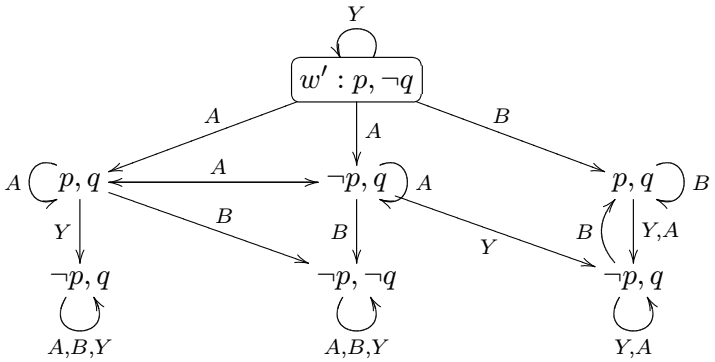

Figure 3. From possible world to multi-agent possible world

all the results of this paper (including the axiomatization) would still hold if we added a common belief operator to the language.

Example 3.3. We see in Figure 3 that a multi-agent possible world is really a generalization of a possible world.

In the single agent case (in AGM belief revision theory), the epistemic state of the agent Y is represented by a finite set of possible worlds which expresses the fact that the agent might have some uncertainty about the situation. In a multi-agent setting, this is very similar: the epistemic state of the agent Y is represented by a (disjoint and) finite set of *multi-agent* possible worlds.

Definition 3.4. An *internal model* is a disjoint and finite union of multi-agent possible worlds.

An internal model will sometimes be noted $(\mathcal{M}, W_a)$ where W_a are the roots of its multi-agent possible worlds.

Example 3.5. In Figure 4 is depicted an example of internal model. This internal model represents how the situation described in Example 2.4 is

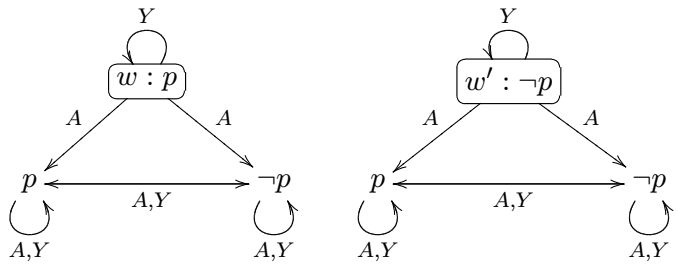

Figure 4. An internal model

perceived by Yann. Yann does not know wether the coin is heads or tails up (formally $\neg B_Y p \wedge \neg B_Y \neg p$). Indeed, in one multi-agent possible world (on the left) p is true at the root and in the other (on the right) p is false at the root. Yann also believes that Alice does not know whether p is true or false (formally $B_Y(\neg B_A p \wedge \neg B_A \neg p)$). Indeed, in both multi-agent possible worlds, $\neg B_A p \wedge \neg B_A \neg p$ is true at the roots. Finally, Yann believes that Ann believes that he does not know whether p is true or false (formally $B_Y B_A(\neg B_Y p \wedge \neg B_Y \neg p)$) since $B_A(\neg B_Y p \wedge \neg B_Y \neg p)$ is true at the roots of both multi-agent possible worlds.

As we said in the introduction, the internal approach can be applied in artificial intelligence. In this case, agent Y is an artificial agent (such as a robot) that has an internal model 'in her mind'. But to stick with a more standard approach (used in the single agent case), we could perfectly consider that agent Y has sentences from a particular language 'in her mind' and draws inferences from them. In that respect, this language could also be used by agent Y in the former approach to perform some model checking in her internal model in order to reason about the situation or to answer queries. So in any case we do need to define a language.

3.2. Language for the internal approach

The well-formed formulas of the language for the internal approach are identical to the ones of the epistemic language of Definition 2.2. Its truth conditions are nevertheless a bit different and are set out below.

Definition 3.6. Let $(\mathcal{M}, \{w^1, \ldots, w^n\}) = \{(M^1, w^1), \ldots, (M^n, w^n)\}$ be an internal model and let $w \in \mathcal{M}$. Then $w \in M^k$ for some k, with $M^k =$

(W^k, R^k, V^k). $\mathcal{M}, w \models \phi$ is defined inductively as follows:

$$
\begin{array}{lll}
\mathcal{M}, w \models \top & & \\
\mathcal{M}, w \models p & \text{iff} & w \in V^k(p) \\
\mathcal{M}, w \models \neg\phi & \text{iff} & \text{not } \mathcal{M}, w \models \phi \\
\mathcal{M}, w \models \phi \wedge \phi' & \text{iff} & \mathcal{M}, w \models \phi \text{ and } \mathcal{M}, w \models \phi' \\
\mathcal{M}, w \models B_Y\phi & \text{iff} & \left\{ \begin{array}{ll} \text{for all } w^i \in W_a,\ \mathcal{M}, w^i \models \phi & \text{if } w \in W_a \\ \text{for all } w' \in R^k_Y(w),\ \mathcal{M}, w' \models \phi & \text{if } w \notin W_a \end{array} \right. \\
\mathcal{M}, w \models B_j\phi & \text{iff} & \text{for all } w' \in R^k_j(w),\ \mathcal{M}, w' \models \phi \qquad \text{if } j \neq Y
\end{array}
$$

We say that ϕ is *true* in $(\mathcal{M}, W_a)$ and write $\mathcal{M}, W_a \models \phi$ when $\mathcal{M}, w \models \phi$ for all $w \in W_a$.

Note that the truth condition for the operator B_Y is defined as if there were accessibility relations indexed by Y between the roots of the multi-agent possible worlds. Therefore we could actually set an accessibility relation indexed by Y between the roots of an internal model. This would lead us to define the notion of *internal model of type 2*, which is a multi-pointed epistemic model $(\mathcal{M}, W_a)$ whose accessibility relations are serial, transitive and euclidean and such that for all $w_a \in W_a$, called the *actual equivalence class*, $R_Y(w_a) = W_a$. One can easily show that the notions of internal model and internal model of type 2 are actually equivalent.[‡] Nevertheless, we prefer to stick to our current representation of internal models because it is more appropriate in the dynamic setting of autonomous agents. Indeed, an important feature of multi-agent possible worlds warranted by condition 2 is their modularity: the agents j's beliefs about agent Y's beliefs (with $j \neq Y$) of a given multi-agent possible world stay the same whatever other multi-agent possible world we add to this multi-agent possible world. This modularity, which is the main motivation for introducing condition 2 in Definition 3.1, plays an important role in the generalization of AGM belief revision theory to the multi-agent case [3, 4].

Thanks to the truth conditions we can now define the notions of satisfiability and validity of a formula. The truth conditions are defined for any world of the internal model. However, the satisfiability and the validity should not be defined relatively to any possible world of the internal model (as it is usually done in epistemic logic) but only to the possible worlds of the actual equivalence class. Indeed, these are the worlds that do count for

[‡]The notion of equivalence between an internal model $(\mathcal{M}, W_a)$ and an internal model of type 2 $(\mathcal{M}', W'_a)$ can be defined naturally by stating that for every $w \in W_a$ there is a $w' \in W'_a$ which satisfies the same formulas as w, and vice versa.

agent Y in an internal model: they are the worlds that agent Y actually considers possible. The other possible worlds are just here for technical reasons in order to express the other agents' beliefs (in these worlds). This leads us to the following definition of satisfiability and validity.

DEFINITION 3.7. Let $\phi \in \mathcal{L}$. The formula ϕ is *internally satisfiable* if there is an internal model $(\mathcal{M}, W_a)$ and there is $w \in W_a$ such that $\mathcal{M}, w \models \phi$. The formula ϕ is *internally valid* if for all internal models $(\mathcal{M}, W_a)$, $\mathcal{M}, W_a \models \phi$. In this last case we write $\models_{\mathsf{Int}} \phi$.

REMARK 3.8. One could define the notions of internal satisfiability and internal validity differently. One could say that $\phi \in \mathcal{L}$ is satisfiable if there is an internal model $(\mathcal{M}, W_a)$ such that $\mathcal{M}, W_a \models \phi$. Then, following this new definition, $\phi \in \mathcal{L}$ is valid if for every internal model $(\mathcal{M}, W_a)$, there is $w \in W_a$ such that $\mathcal{M}, w \models \phi$.

This second notion of internal validity corresponds to Gärdenfors' notion of validity [15]. These two notions of internal validity also correspond to the two notions of validity studied by Levi and Arlo Costa [2]: they call the first one "positive validity" and the second one "negative validity".

These two notions coincide if we use a propositional language but not if we use an epistemic one (and therefore in particular not in a multi-agent setting). Indeed, the Moore sentence $p \wedge \neg B_Y p$ is positively satisfiable but not negatively satisfiable. Nevertheless there are some connections between them. We can indeed prove that $\phi \in \mathcal{L}$ is positively valid if and only if $B_Y \phi$ is negatively valid. Moreover, both have advantages and drawbacks. On the one hand, positive validity is intuitive because it says that a formula ϕ is valid if in every possible situation, the agent Y believes ϕ. However positive satisfiability is less intuitive because ϕ is positively satisfiable if there exists a situation in which the agent Y does not reject ϕ. On the other hand, negative satisfiability is also intuitive because it says that ϕ is negatively satisfiable if there exists a situation in which agent Y believes ϕ. However negative validity is less intuitive because it says that ϕ is negatively valid if in every situation agent Y does not reject ϕ.

In modal logic [8] there are two notions of semantic consequence. In the internal approach we can also define two notions of semantic consequence. Let C be a class of internal models, $\Sigma \subseteq \mathcal{L}$ and $\phi \in \mathcal{L}$. First, we say that ϕ is a *local internal consequence* of Σ over C, written $\Sigma \models_{\mathsf{C}} \phi$, if for all internal models $(\mathcal{M}, W_a) \in \mathsf{C}$ and all $w \in W_a$, if $\mathcal{M}, w \models \Sigma$ then $\mathcal{M}, w \models \phi$. Second, we say that ϕ is a *global internal consequence* of Σ over C, written $\Sigma \models^g_{\mathsf{C}} \phi$, if for all internal models $(\mathcal{M}, W_a) \in \mathsf{C}$, if $\mathcal{M}, W_a \models \Sigma$ then $\mathcal{M}, W_a \models \phi$. For

example, if we take any class C of internal models then it is not necessarily the case that $\phi \models_{\mathsf{C}} B_Y \phi$, whereas we do have that $\phi \models^g_{\mathsf{C}} B_Y \phi$. Local internal consequence can be associated to positive satisfiability and global internal consequence can be associated to negative satisfiability of the above Remark.

4. Some connections between the internal and the external approaches

4.1. Internal approach and *perfect* external approach

Intuitively, there are some connections between the internal and the perfect external approaches. Indeed, in the perfect external approach the modeler is supposed to know perfectly how the agents perceive the surrounding world. So from the model she builds we should be able to extract the internal model of each agent. Likewise, it seems natural to claim that for the agent Y a formula is true if and only if, objectively speaking, the agent Y believes this formula. In this section we are going to formalize these intuitions. First we define the notion of perfect external model.

Definition 4.1. A *perfect external model* (for the agents G) is a pointed epistemic model $(M, w_a) = (W, R, V, w_a)$ where $w_a \in W$ and the accessibility relations R_j are defined for $j \in G$, serial, transitive and euclidean.

So what we call a perfect external model is just a standard pointed epistemic model used in epistemic logic. A perfect external model is supposed to model truthfully and from an external point of view how all the agents involved in the same situation perceive the actual world (represented formally by w_a). This is thus simply the type of model built by the modeler in the perfect external approach spelled out in the introduction. The language and truth conditions for these perfect external models are the same as in epistemic logic and are spelled out in Definitions 2.2 and 2.3. The notion of perfect external validity is also the same as in epistemic logic and we say that $\phi \in \mathcal{L}$ is *perfectly externally valid*, noted $\models_{\mathsf{Ext}} \phi$, if for all perfect external model (M, w), $M, w \models \phi$ (and similarly for *perfect external satisfiability*).

Proposition 4.2. *Let (M, w_a) be a perfect external model (for the agents G) and let $j \in G$. The* model associated to agent j and (M, w_a) *is the submodel of M generated by $R_j(w_a)$. It is an internal model (of type 2) and $R_j(w_a)$ is its actual equivalence class.*

Because the perfect external model is supposed to model truthfully the situation, w_a does correspond formally to the actual world. So $R_j(w_a)$ are

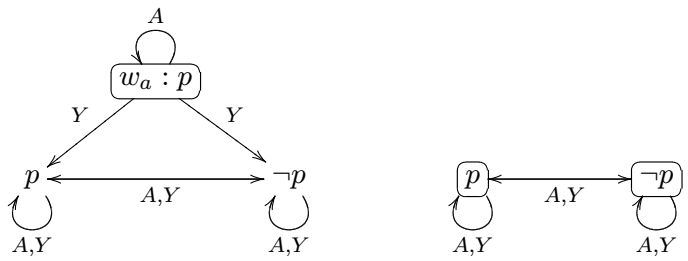

Figure 5. External model (M, w_a) (*left*); Internal model associated to Yann (*right*), Internal model associated to Alice (*left*).

the worlds that agent j actually considers possible in this situation. In agent j's internal model pertaining to this situation, these worlds should then be the worlds of the actual equivalence class. Finally, taking the submodel generated by these worlds ensures that the piece of information encoded in the worlds $R_j(w_a)$ in the perfect external model is kept unchanged in the associated internal model.

Example 4.3. In Figure 5 is depicted the 'coin example' after Alice's cheating (see introduction). We can check that in the perfect external model, Yann does not know whether the coin is heads or tails up and moreover believes that Alice does not know neither. This is also true in the internal model (of type 2) associated to Yann. However, in the perfect external model, Alice knows that the coin is heads up but this is false in the internal model associated to Yann and true in Alice's internal model. Note that the internal model associated to Yann is actually equivalent to the internal model of Figure 4. Note also that the internal model associated to Alice is the same as the perfect external model. This is due to the fact that Alice perceived correctly what happened in the situation.

So we know from a perfect external model how to obtain the internal model of each agent. But conversely, one obviously cannot obtain from the internal model of one of the agents the perfect external model representing the situation, which is in line with our intuitions. We can nevertheless get the perfect external model if we suppose given the internal models of *all* the agents and if we assume that the modeler knows the real state of the world, more precisely she knows what propositional facts are true in the actual world. We can then introduce a single world w_a whose valuation V_a satisfies these propositional facts. The perfect external model is built by setting accessibility relations indexed by j from w_a to the actual equivalence

class of j's internal model (of type 2), and so for each agent $j \in G$.

As we said in Section 3.2, the language of the internal approach is the same as that of the perfect external approach. This enables us to draw easily some connections between the two approaches.

PROPOSITION 4.4. *For all* $\phi \in \mathcal{L}$, $\models_{\mathsf{Int}} \phi$ *iff* $\models_{\mathsf{Ext}} B_Y \phi$.

Intuitively, this result is correct: for you ϕ is true if and only if from an external point of view you believe that ϕ is the case. (Note that this result does not hold for the notion of negative validity.)

4.2. Internal approach and *imperfect* external approach

In both the internal approach and the imperfect external approach the modeler might have some uncertainty about the situation to model (unlike the perfect external approach). On the other hand, unlike the imperfect external approach, in the internal approach the modeler is one of the agents G under consideration and is therefore present in the formalism as agent $Y \in G$. So in the imperfect internal approach and for this uncertain and external modeler a *possible* representation of the situation is simply a perfect external model. This leads us to define the notion of imperfect external model.

DEFINITION 4.5. An *imperfect external model* (for the agents G) is a finite set of perfect external models (for the agents G). [§]

REMARK 4.6. Sometimes in the literature, the uncertainty of the modeler is taken to be a property of the object of study itself. For example, in [17], some events are deemed indeterministic whereas their 'indeterminism' is actually due to a lack of knowledge about the events from the part of the modeler. To avoid these kinds of confusion, one should instead clearly specify from the start the modeling approach and the object of study.

There are obviously connections between the internal and the imperfect external approaches. Indeed, as we just said the modeler-agent Y is present in the internal model representing the situation. However, she could also 'step back' and model the situation from an external point of view as if she was not present. In that case we would shift from an internal approach to an imperfect external approach because the modeler has still some uncertainty about the situation. From a semantic point of view, this change of

[§]Equivalently, an imperfect external model can be defined as a multi-pointed epistemic model.

perspective simply amounts to suppress from an internal model $(\mathcal{M}, W_a)$ the accessibility relations indexed by Y. This yields an imperfect external model (for the agents $G - \{Y\}$) $\mathsf{Imp}(\mathcal{M}, W_a)$. From a syntactic point of view, this change of perspective somehow amounts to 'forget' agent Y from formulas. There are obviously several ways to define from a formula ϕ of $\mathcal{L}$ a formula $\mathsf{Imp}(\phi)$ of the same language $\mathcal{L}$ but without B_Y operator. We give in the appendix a proposal (inspired from the forget operator for first-order logic originally introduced in [22]) such that if ϕ is true in the internal model $(\mathcal{M}, W_a)$ then $\mathsf{Imp}(\phi)$ is true in the imperfect external model $\mathsf{Imp}(\mathcal{M}, W_a)$, and such that $\mathsf{Imp}(\phi)$ functions as a prime implicate of ϕ. Finally, the modeler in the imperfect external approach can really be seen as the modeler-agent Y of the internal approach who would model the situation as if she was not present. Indeed, one can show that for any imperfect external model (for the agents $G-\{Y\}$) $(\mathcal{M}', W'_a)$ there is an internal model $(\mathcal{M}, W_a)$ such that $\mathsf{Imp}(\mathcal{M}, W_a) = (\mathcal{M}', W'_a)$.

5. Axiomatization of the internal semantics

As we said earlier, instead of internal models, agent Y might have formulas 'in her mind' in order to represent the surrounding world. But to draw inferences from them she needs a proof system. In other words, we still need to axiomatize the internal semantics. That is what we are going to do now.

First some notations. Let Ext designate from now on the logic $\mathsf{KD45_G}$. So for all $\phi \in \mathcal{L}$, $\vdash_{\mathsf{Ext}} \phi$ iff $\phi \in \mathsf{KD45_G}$.

Definition 5.1. The *internal logic* Int is defined by the following axiom schemes and inference rules:

T	$\vdash_{\mathsf{Int}} B_Y\phi \to \phi$;
I-E	$\vdash_{\mathsf{Int}} \phi$ for all $\phi \in \mathcal{L}$ such that $\vdash_{\mathsf{Ext}} \phi$;
MP	if $\vdash_{\mathsf{Int}} \phi$ and $\vdash_{\mathsf{Int}} \phi \to \psi$ then $\vdash_{\mathsf{Int}} \psi$.

Let us have a closer look at the axiom schemes. The first one tells us that for you, everything you believe is true. This is coherent if we construe the notion of belief as conviction. The second one tells us that you should believe everything which is objectively true, i.e. which is true independently of your own subjectivity. This includes not only propositional tautologies but also the way the other agents reason, like for example the fact that (for you) agent j believes that everything she believes is true (because $\vdash_{\mathsf{Ext}} B_j(B_j\phi \to \phi)$). Finally note that the necessitation rule ($\vdash_{\mathsf{Int}} \phi$ implies $\vdash_{\mathsf{Int}} B_j\phi$ for all j) is not present, which is intuitively correct. Indeed, if for you ϕ is true (i.e. you

believe ϕ) then in general there is no reason that you should believe that the other agents believe ϕ as well. For example, $B_Y\phi \rightarrow \phi$ is internally valid but $B_j(B_Y\phi \rightarrow \phi)$ (for $j \neq Y$) should not be internally valid.

THEOREM 5.2 (soundness and completeness). *For all $\phi \in \mathcal{L}$,*

$$\models_{\mathit{Int}} \phi \text{ iff } \vdash_{\mathit{Int}} \phi.$$

PROOF SKETCH. Soundness being straightforward, we only focus on completeness. Let ϕ be a Int-consistent formula. We need to prove that there is an internal model $(\mathcal{M}_{\mathsf{Int}}, W_a)$, there is $w \in W_a$ such that $\mathcal{M}_{\mathsf{Int}}, w \models \phi$.

The model $\mathcal{M}_{\mathsf{Int}}$ is built as follows. Let $Sub^+(\phi)$ be all the subformulas of ϕ with their negations. Let W_{Int} be the set of maximal Int-consistent subsets of $Sub^+(\phi)$. Let W_{Ext} be the set of maximal Ext-consistent subsets of $Sub^+(\phi)$. For all $\Gamma, \Gamma' \in W_{\mathsf{Int}} \cup W_{\mathsf{Ext}}$, let $\Gamma/B_j = \{\psi; B_j\psi \in \Gamma\}$ and $B_j\Gamma = \{B_j\psi; B_j\psi \in \Gamma\} \cup \{\neg B_j\psi; \neg B_j\psi \in \Gamma\}$. We define the epistemic model $M = (W, R, V)$ as follows: $W = W_{\mathsf{Int}} \cup W_{\mathsf{Ext}}$; for all $j \in G$ and $\Gamma, \Gamma' \in W$, $\Gamma' \in R_j(\Gamma)$ iff $\Gamma/B_j = \Gamma'/B_j$ and $\Gamma/B_j \subseteq \Gamma'$; $\Gamma \in V(p)$ iff $p \in \Gamma$.

One then proves by induction on ψ that for all $\psi \in Sub^+(\phi)$, all $\Gamma \in W$, $M, \Gamma \models \psi$ iff $\psi \in \Gamma$. Then one can show that the accessibility relations R_j are serial, transitive and euclidean and that for all $\Gamma \in W_{\mathsf{Int}}$, $\Gamma \in R_Y(\Gamma)$ $(*)$.

ϕ is a Int-consistent formula so there is $\Gamma \in W_{\mathsf{Int}}$ such that $\phi \in \Gamma$, i.e. $M, \Gamma \models \phi$. Let $\mathcal{M}_{\mathsf{Int}}$ be the submodel generated by $R_Y(\Gamma)$. Then clearly $(\mathcal{M}, W_a)$ with $W_a = R_Y(\Gamma)$ is an internal model. Finally, because $\Gamma \in R_Y(\Gamma)$ by $(*)$, there is $\Gamma \in W_a$ such that $\mathcal{M}_{\mathsf{Int}}, \Gamma \models \phi$. ■

From this axiomatization we can also prove that for all $\phi \in \mathcal{L}$, $\vdash_{\mathsf{Int}} \phi$ iff $\vdash_{\mathsf{Ext}} B_Y\phi$ and $\vdash_{\mathsf{Int}} \phi$ iff $\vdash_{\mathsf{Int}} B_Y\phi$. This entails that in case Y is the only agent then the internal logic Int is actually the logic $\mathsf{S5}$, which is also the logic advocated by Isaac Levi as the logic of 'full belief' (for the internal approach) [21].

Finally the internal logic Int has also nice computational properties. Its complexity turns out to be the same as in the perfect external approach.

THEOREM 5.3. *The validity problem for the internal logic Int is decidable and PSPACE-complete.*

REMARK 5.4. Soon after Hintikka's seminal book was published [18], an issue now known as the logical omniscience problem was raised by Castañeda about Hintikka's epistemic logic: his "senses of 'knowledge' and 'belief' are

much too strong [...] since most people do not even understand all deductions from premises they know to be true" [10]. It sparked a lot of work aimed at avoiding this problem (such as [20], [13] or [12]).

While we believe that it is indeed a problem when we want to model or describe human-like agents, we nevertheless believe that it is not really a problem when we want to design artificial agents. Indeed, these agents are supposed to reason according to our internal logic and because of its decidability, artificial agents are in fact logically omniscient (even if it will take them some time to compute all the deductions given the complexity of our logic).

6. Conclusion

In the introduction, we have identified what we claim to be the only three logically possible modeling approaches by proceeding by successive dichotomies. Such a classification has not been made before. Afterwards, we have focused on the internal approach for which a logical formalism in a multi-agent setting is missing, although such a formalism is crucial if we want to design autonomous agents for instance. We have proposed one by generalizing the possible world semantics of the AGM belief revision theory. This formalism enabled us to draw some formal links between the external and the internal approaches which are in line with our intuitions of these approaches. Finally, we have provided an axiomatization of our formalism whose axioms are also in line with our intuitions of the internal approach and showed that our logic is decidable and PSPACE-complete.

In this paper we stressed the importance of specifying a modeling point of view and an object of study when one wants to model epistemic situations. As it turns out, even if there are important conceptual and intuitive differences between the internal and external approaches, the corresponding formalisms are rather close and can be easily mapped one to the other. However, it is this formally little but conceptually important difference which allows to easily introduce dynamics in the internal approach and to generalise the AGM results on belief revision to the multi-agent case, as we show in [3, 4].

We finally sketch how our internal version of epistemic logic allows to easily lift AGM belief revision theory to the multi-agent case (see [3, 4] for more details). Following the belief base approach we represent a belief base in a multi-agent setting by an epistemic formula ψ. Then we replace possible worlds in AGM theory by *multi-agent* possible worlds and we replace the propositional language of AGM theory by the epistemic language. This

means that the models of the epistemic formula ψ are the multi-agent possible worlds that satisfy ψ. Then the definition of a faithful ordering $\leq_\psi$ on multi-agent possible worlds for a given epistemic formula ψ is the same as the definition of a faithful ordering $\leq_{\psi'}$ on possible worlds for a given propositional formula ψ'. Intuitively, $(M, w) \leq_\psi (M', w')$ means that the multi-agent possible world (M, w) is closer to ψ than the multi-agent possible world (M', w'). Likewise, the rationality postulates for belief revision in a multi-agent setting are the same as in the AGM theory except that propositional formulas are replaced by epistemic formulas. Then we can show, as in the AGM theory, that a revision operator satisfies these postulates in a multi-agent setting if and only if the models of the revision of the belief base ψ by the epistemic formula ϕ are the multi-agent possible worlds that satisfy ϕ and which are minimal with respect to $\leq_\psi$.

Acknowledgements. I thank Isaac Levi and my PhD supervisors Andreas Herzig and Hans van Ditmarsch for interesting and useful discussions on the topic of this paper.

References

[1] Alchourrón, Carlos, Peter Gärdenfors, and David Makinson, 'On the logic of theory change: Partial meet contraction and revision functions.', *Journal of Symbolic Logic*, 50 (1985), 2, 510–530.

[2] Arló Costa, Horacio, and Isaac Levi, 'Two notions of epistemic validity (epistemic models for Ramsey's conditionals)', *Synthese*, 109 (1996), 217–262.

[3] Aucher, Guillaume, 'Internal models and private multi-agent belief revision', in Muller Padgham, Parkes, and Parsons, (eds.), *Proceedings of Autonomous Agents and Multi-agent Systems (AAMAS 2008)*, Estoril, Portugal, 12-16 May 2008, pp. 721–727.

[4] Aucher, Guillaume, 'Generalizing AGM to a multi-agent setting', *Logic Journal of the IGPL*, (2009). To appear.

[5] Baltag, Alexandru, and Larry Moss, 'Logic for epistemic programs', *Synthese*, 139 (2004), 2, 165–224.

[6] Banerjee, Mohua, and Didier dubois, 'A simple modal logic for reasoning about revealed beliefs', in Claudio Sossai, and Gaetano Chemello, (eds.), *ECSQARU*, vol. 5590 of *LNCS*, Springer, 2009, pp. 805–816.

[7] Battigalli, Pierpaolo, and Giacomo Bonanno, 'Recent results on belief, knowledge and the epistemic foundations of game theory', *Research in Economics*, 53 (1999), 149–225.

[8] Blackburn, Patrick, Maarten de Rijke, and Yde Venema, *Modal Logic*, vol. 53 of *Cambridge Tracts in Computer Science*, Cambridge University Press, 2001.

[9] Booth, Richard, and Alexander Nittka, 'Reconstructing an agent's epistemic state from observations about its beliefs and non-beliefs', *Journal of Logic and Computation*, (2007). Accepted for publication.

[10] Castañeda, Hector-Neri, 'Review of 'knowledge and belief'', *Journal of Symbolic Logic*, 29 (1964), 132–134.

[11] Cohen, Philip, and Hector Levesque, 'Intention is choice with commitment', *Artificial intelligence*, 42 (1990), 213–261.

[12] Duc, Ho Ngoc, *Resource-Bounded Reasoning about Knowledge*, Ph.D. thesis, University of Leipzig, 2001.

[13] Fagin, Ronald, and Joseph Halpern, 'Belief, awareness, and limited reasoning', *Artificial Intelligence*, 34 (1988), 39–76.

[14] Fagin, Ronald, Joseph Halpern, Yoram Moses, and Moshe Vardi, *Reasoning about knowledge*, MIT Press, 1995.

[15] Gärdenfors, Peter, *Knowledge in Flux (Modeling the Dynamics of Epistemic States)*, Bradford/MIT Press, Cambridge, Massachusetts, 1988.

[16] Georgeff, Michael, and Anand Rao, 'Asymmetry thesis and side-effect problems in linear time and branching time intention logics', in *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI-91)*, (Sydney, Australia), 1991, pp. 498–504.

[17] Herzig, Andreas, Jérôme Lang, and Pierre Marquis, 'Revision and update in multiagent belief structures', in *5th Conference on Logic and the Foundations of Game and Decision Theory (LOFT6)*, Leipzig, 2004.

[18] Hintikka, Jaakko, *Knowledge and Belief, An Introduction to the Logic of the Two Notions*, Cornell University Press, Ithaca and London, 1962.

[19] Lenzen, Wolfgang, *Recent Work in Epistemic Logic*, Acta Philosophica 30, North Holland Publishing Company, 1978.

[20] Levesque, Hector, 'A logic of implicit and explicit knowledge', in *AAAI-84*, Austin Texas, 1984, pp. 198–202.

[21] Levi, Isaac, *The covenant of reason: rationality and the commitments of thought*, Cambridge University Press, 1997.

[22] Lin, Fangzhen, and Ray Reiter, 'Forget it!', in *Proceedings of the AAAI Fall Symposium on Relevance*, 1994, pp. 154–159.

[23] Meyer, John-Jules Ch., Frank de Boer, Rogier van Eijk, Koen Hindriks, and Wiebe van der Hoek, 'On programming KARO agents', *Logic Journal of the IGPL*, 9 (2001), 2.

[24] Nagel, Thomas, *The view from nowhere*, oxford university press, 1986.

[25] Nittka, Alexander, *A Method for Reasoning about other Agents' Beliefs from Observations*, Ph.D. thesis, University of Leipzig, 2008.

[26] Rao, Anand, and Michael Georgeff, 'Modeling rational agents within a BDI-architecture', in R. Fikes, and E. Sandewall, (eds.), *Proceedings of Knowledge Representation and Reasoning (KR & R-91)*, Morgan Kaufmann Publishers, 1991, pp. 473–484.

[27] van Linder, Bernd, Wiebe van der Hoek, and John-Jules Ch. Meyer, 'Formalising abilities and opportunities of agents', *Fundamenta Informaticae*, 34 (1998), 1-2, 53–101.

[28] Voorbraak, Frans, *As Far as I know. Epistemic Logic and Uncertainty*, Ph.D. thesis, Utrecht University, 1993.
[29] Wooldridge, Michael, *Reasoning About Rational Agents*, MIT Press, 2000.

A. Forgetting agents

We define an operator Imp on formulas which somehow 'forgets' agent Y from formulas. Let $\phi \in \mathcal{L}$ and let $\phi_1, \ldots, \phi_n$ be the subformulas of ϕ of the form $B_Y\psi$ not within the scope of another operator B_Y. We define $\mathsf{Imp}(\phi) = \mathsf{Imp}(\phi, \{\phi_1, \ldots, \phi_n\}) \in \mathcal{L}_{G-Y}$ inductively as follows: $\mathsf{Imp}(\phi, \{\psi\}) = \phi[\psi/\top] \vee \phi[\psi/\bot]$ (where $\phi[\psi/\top]$, resp. $\phi[\psi/\bot]$, is the formula ϕ where ψ has been substituted by $\top$, resp. by $\bot$) and $\mathsf{Imp}(\phi, \{\psi\} \cup S) = \mathsf{Imp}(\mathsf{Imp}(\phi, \{\psi\}), S)$. We then have the following result.

Proposition A.1. *For all $\phi \in \mathcal{L}$ and $\phi' \in \mathcal{L}_{G-Y}$,*

1. $\models_{\mathsf{Ext}} \phi \to \mathsf{Imp}(\phi)$;
2. *If* $\models_{\mathsf{Ext}} \phi \to \phi'$ *then* $\models_{\mathsf{Ext}} \mathsf{Imp}(\phi) \to \phi'$
3. $(\mathcal{M}, W_a) \models \phi'$ *iff* $\mathsf{Imp}(\mathcal{M}, W_a) \models \phi'$.[¶]

Proof Sketch. For item 1, we prove by induction on the degree of $\phi \in \mathcal{L}$ that there is $\chi_\phi \in \mathcal{L}$ such that $\models_{\mathsf{Ext}} (\chi_\phi \to (\phi \leftrightarrow \phi[B_Y\psi/\top])) \wedge (\neg\chi_\phi \to (\phi \leftrightarrow \phi[B_Y\psi/\bot]))$. We then get the result using $\models_{\mathsf{Ext}} \chi_\phi \vee \neg\chi_\phi$ and modus ponens. For item 2, if (M, w) is a perfect external model for the set of agents $G - \{Y\}$ such that $M, w \models \phi[B_Y\psi/\top] \vee \phi[B_Y\psi/\bot]$, we can easily build a perfect external model (M', w') for the set of agents G satisfying the same formulas of $\mathcal{L}_{G-Y}$ as (M, w) and such that $M', w' \models \phi$. Therefore $M', w' \models \phi'$ by assumption, and so $M, w \models \phi'$ because $\phi' \in \mathcal{L}_{G-Y}$. ∎

The first two items are valid in the external approach but also hold if we use instead the internal validity. The first one intuitively expresses that our forgetting operation Imp is sound. The second one intuitively expresses that $\mathsf{Imp}(\phi)$ is the formula of $\mathcal{L}_{G-Y}$ derived from ϕ which retains from ϕ the most important amount of information possible about the agents G different from Y (therefore functioning like a prime implicate).

[¶] $\mathsf{Imp}(\mathcal{M}, W_a) \models \phi$ means that ϕ is true at the root of each perfect external model of $\mathsf{Imp}((M), W_a)$.

Guillaume Aucher
Université Paul Sabatier (F) – University of Otago (NZ)
IRIT – Équipe LILaC
118, route de Narbonne
F-31062 Toulouse cedex 9 (France)
aucher@irit.fr