Internal Models and Private Multi-agent Belief Revision

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

ABSTRACT

We generalize AGM belief revision theory to the multi-agent case. To do so, we first generalize the semantics of the single-agent case, based on the notion of interpretation, to the multi-agent case. Then we show that, thanks to the shape of our new semantics, all the results of the AGM framework transfer. Afterwards we investigate some postulates that are specific to our multi-agent setting.

Categories and Subject Descriptors

I.2.4. [**Knowledge Representation Formalism and Methods**]: Modal Logic; I.2.1.1 [**Distributed Artificial Intelligence**]: Intelligent Agents, Multi-agent systems; I.2.3 [**Deduction and Theorem Proving**]: non-monotonic reasoning and belief revision

General Terms

Theory

Keywords

Belief revision, Epistemic logic, Multi-agent systems

1. INTRODUCTION

AGM belief revision theory [1] has been designed for a single agent. It seems natural to extend it to the multi-agent case. As in AGM, we consider the beliefs of *one* agent, that we call Y (like You). But in this case this agent, in her representation of the surrounding world, will have to deal not only with facts about the world but also with how the other agents perceive the surrounding world. So, we will have to extend or generalize the single agent semantics in order to take into account this multi-agent aspect. Such a formalism is crucial if we want to design autonomous agents able to act in a multi-agent setting.

Besides, in a multi-agent setting, we have to be careful about what kind of multi-agent belief revision we study and consequently about the nature of the events we consider. In this paper we are interested in *private announcements* made to Y. A private announcement is an event where Y learns privately (from an external source for example) some piece of information about the original situation, the other agents not being aware of anything. This piece of information might be factual or epistemic, i.e. about some other agents' beliefs. Finally, by private multi-agent belief *revision*, we mean the revision that Y must perform in case the private announcement of φ made to her contradicts her beliefs. So far, this kind of revision has not been studied.

In the case of private announcement, the other agents' belief clearly do not change. For example, suppose $y$ou (Y) believe p, and agent j believes p (and perhaps even that p is common belief of Y and j). When a third external agent privately tells you that $\neg p$ then j still believes p and you still believe that j believes p (and that j believes that p is common belief). This static aspect of private announcements is similar to the static aspect of AGM belief revision in a single-agent case: in both cases the world does not change but only agent Y's beliefs about the world change. So, it is reasonable to expect that the AGM framework can be extended to private multi-agent belief revision. In this paper we propose a natural generalization. The central device will be internal models, of which AGM models are a particular case.

The paper is organized as follows. In Section 2, we recall belief revision theory in the line of [8]. In Section 3, we first introduce the notions of multi-agent possible worlds and internal models in order to adequately represent agent Y's perception of the surrounding world. We also propose an equivalent representation. Then we generalize the results of Section 2 to the multi-agent case. Finally in Section 4, we investigate some additional rationality postulates specific to our multi-agent approach.

2. THE SINGLE AGENT CASE: THE AGM APPROACH

In this paper Φ is a *finite* set of propositional letters and L the propositional language defined over Φ. Often, the epistemic state of the agent at stake is represented by a belief set K. This belief set is an infinite set of propositional formulas closed under logical consequence and whose formulas represent the beliefs of the agent. However, we prefer to represent epistemic states by finite belief base as it is easier to handle by computers. For that, we follow the approach of [8].

As argued by Katsuno and Mendelzon, because Φ is finite, a belief set K can be equivalently represented by a mere propositional formula ψ: $K = Cn(\psi) = \{\varphi; \psi \rightarrow \varphi\}$. So $\varphi \in K$ iff $\psi \rightarrow \varphi$. Now, given a belief base ψ and a sentence μ, $\psi \circ \mu$ denotes the revision of ψ by μ; that is the new

Cite as: Internal Models and Private Multi-agent Belief Revision, Guillaume Aucher, *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, Padgham, Parkes, Müller and Parsons (eds.),May,12-16.,2008,Estoril,Portugal,pp. 721-727.

belief base obtained by adding μ to the old belief base ψ and giving up some formulas if necessary to keep consistency. In fact, given a revision operator $*$ on belief sets, one can define a corresponding operator $\circ$ on belief bases as follows: $\psi \circ \mu \rightarrow \varphi$ iff $\varphi \in Cn(\psi) * \mu$. Thanks to this correspondence, Katsuno and Mendelzon set some rationality postulates for this revision operator $\circ$ on belief bases which are equivalent to the AGM rationality postulates for the revision $*$ on belief sets. These postulates express how a rational agent should revise her belief set when she receives incoming information that she believes to be true.

LEMMA 1. *[8] Let * be a revision operator on belief sets and $\circ$ its corresponding operator on belief bases. Then * satisfies the 8 AGM postulates $(K*1)-(K*8)$ iff $\circ$ satisfies the postulates $(R1)-(R6)$ below:*

(R1) $\psi \circ \mu \rightarrow \mu$.

(R2) *if $\psi \wedge \mu$ is satisfiable, then $\psi \circ \mu \leftrightarrow \psi \wedge \mu$.*

(R3) *If μ is satisfiable, then $\psi \circ \mu$ is also satisfiable.*

(R4) *If $\psi_1 \leftrightarrow \psi_2$ and $\mu_1 \leftrightarrow \mu_2$, then $\psi_1 \circ \mu_1 \leftrightarrow \psi_2 \circ \mu_2$.*

(R5) $(\psi \circ \mu) \wedge \varphi \rightarrow \psi \circ (\mu \wedge \varphi)$.

(R6) *If $(\psi \circ \mu) \wedge \varphi$ is satisfiable, then $\psi \circ (\mu \wedge \varphi) \rightarrow (\psi \circ \mu) \wedge \varphi$.*

So far our approach to revision was syntactically driven. Now we are going to give a semantical approach to revision and then set some links between the two approaches.

Let $\mathcal{I}$ be the set of all the interpretations of the finite propositional language L. $Mod(\psi)$ denotes the set of all the interpretations that make ψ true. Let $\mathcal{M}$ be a set of interpretations of L. $form(\mathcal{M})$ denotes a formula whose set of models is equal to $\mathcal{M}$.

A *pre-order* $\leq$ over $\mathcal{I}$ is a reflexive and transitive relation on $\mathcal{I}$. A pre-order is *total* if for every $I, J \in \mathcal{I}$, either $I \leq J$ or $J \leq I$. Consider a function that assigns to each propositional formula ψ a pre-order $\leq_\psi$ over $\mathcal{I}$. We say this assignment is *faithful* if the following three conditions hold:

1. If $I, I' \in Mod(\psi)$, then $I <_\psi I'$ does not hold.
2. If $I \in Mod(\psi)$ and $I' \notin Mod(\psi)$, then $I <_\psi I'$ holds.
3. If $\psi \leftrightarrow \varphi$, then $\leq_\psi = \leq_\varphi$.

Let $\mathcal{M}$ be a subset of $\mathcal{I}$. An interpretation I is minimal in $\mathcal{M}$ with respect to $\leq_\psi$ if $I \in \mathcal{M}$ and there is no $I' \in \mathcal{M}$ such that $I' <_\psi I$. Let

$$Min(\mathcal{M}, \leq_\psi) := \{I; I \text{ is minimal in } \mathcal{M} \text{ with respect to } \leq_\psi\}$$

THEOREM 1. *[8] Revision operator $\circ$ satisfies postulates $(R1)-(R6)$ iff there exists a faithful assignment that maps each belief base ψ to a total pre-order $\leq_\psi$ such that $Mod(\psi \circ \mu) = Min(Mod(\mu), \leq_\psi)$.*

PROOF. The detailed proof can be found in [8], we just give a sketch of it here. The "if" direction is straightforward. For the "only-if" direction, the key is the definition of a faithful assignment for each belief base in terms of $\circ$. For any interpretations I and I' ($I = I'$is permitted), we define a relation $\leq_\psi$ as $I \leq_\psi I'$ iff either $I \in Mod(\psi)$ or $I \in Mod(\psi \circ form(I, I'))$. $\square$

3. THE MULTI-AGENT CASE

3.1 Some Technical Preliminaries

In this paper, G is a fixed set of agents such that $Y \in G$.

3.1.1 Epistemic logic

We first recall the basics of epistemic logic. An epistemic model M is a tuple $M = (W, \{R_j; j \in G\}, val)$ where W is a set of worlds, R_j are accessibility relations indexed by agents $j \in G$ and val is a function that assigns to each $w \in W$ a subset of Φ. We define $R_j(w)$ by $R_j(w) := \{v; wR_jv\}$ and $|M|$ is the number of worlds in M. Finally, a $KD45_G$ epistemic model is an epistemic model whose accessibility relations are serial, transitive and euclidean.[1]

Classically, an epistemic model M is given with an actual world w_a: (M, w_a). Intuitively, a (pointed) epistemic model (M, w_a) represents from an external point of view how the actual world w_a is perceived by the agents G. The possible worlds W are the relevant worlds needed to define such a representation and the valuation val specifies which propositional letters (such as 'it is raining') are true in these worlds. Finally the accessibility relations R_j models the notion of belief. We set $w' \in R_j(w)$ in case in world w, agent j considers the world w' possible.

Now we can define a language for epistemic models which will enable us to express things about them.

$$\mathcal{L} : \varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \wedge \varphi \mid B_j\varphi \mid C_{G_1}\varphi,$$

where j ranges over G, G_1 over subsets of G and p over Φ. Its semantics is defined as usual as follows.

$$\begin{array}{lll}
M, w \models \top & & \\
M, w \models p & \text{iff} & w \in V(p) \\
M, w \models \neg\varphi & \text{iff} & \text{not } M, w \models \varphi \\
M, w \models \varphi \wedge \varphi' & \text{iff} & M, w \models \varphi \text{ and } M, w \models \varphi' \\
M, w \models B_j\varphi & \text{iff} & \text{for all } v \in R_j(w), M, v \models \varphi \\
M, w \models C_{G_1}\varphi & \text{iff} & \text{for all } v \in (\bigcup_{j \in G_1} R_j)^+(w) M, v \models \varphi
\end{array}$$

where $(\bigcup_{j \in G_1} R_j)^+$ is the transitive closure of $\bigcup_{j \in G_1} R_j$.

$M, w \models B_j\varphi$ intuitively means that in world w, agent j believes that φ is true (because φ is true in all the worlds that the agent j considers possible). For example, in the pointed epistemic model (M, w) of Figure 4, agent Y does not know whether p is true or not: $M, w \models \neg B_Y p \wedge \neg B_Y \neg p$. Agent Y also believes that A does not know neither: $M, w \models B_Y(\neg B_A p \wedge \neg B_A \neg p)$. Finally, agent Y believes that A believes that she does not know whether p is true or not: $M, w \models B_Y B_A(\neg B_Y \neg p \wedge \neg B_Y p)$. $M, w \models C_{G_1}\varphi$ intuitively means that in world w, φ is common belief among the agents G_1, that is to say every agent of G_1 believes φ, and every agent of G_1 believes that every agent of G_1 believes φ... and so on ad infinitum. For example, in the pointed epistemic model (M, w) of Figure 4, it is common belief among all the agents that agent Y does not know whether p is true or not: $M, w \models C_G(\neg B_Y p \wedge \neg B_Y \neg p)$.

[1] An accessibility relation R is

- *serial* if for all w, $R(w) \neq \emptyset$;
- *transitive* if for all w, v, u, if wRv and vRu then wRu;
- *euclidean* if for all w, v, u, if wRv and wRu then vRu.

3.1.2 Bisimulation

We now recall the definition of a bisimulation. Intuitively, two epistemic models are bisimilar if they contain the same information.

DEFINITION 1. *Let Z be a relation between worlds of two finite epistemic models $M = (W, \{R_j; j \in G\}, val)$ and $M' = (W', \{R'_j; j \in G\}, val')$. We define the property of Z being a* bisimulation *in w and w', noted $Z : M, w \leftrightarroweq M', w'$ as follows.*

1. *If wZw' then $val(w) = val'(w')$;*
2. *if wZw' and $v \in R_j(w)$ then there exists $v' \in R_j(w')$ such that vZv';*
3. *if wZw' and $v' \in R_j(w'$ then there exists $v \in R_j(w)$ such that vZv'.*

We can define bisimilarity between M, w and M', w', noted $M, w \leftrightarroweq M', w'$ by $M, w \leftrightarroweq M', w'$ iff there is a relation Z such that $Z : M, w \leftrightarroweq M', w'$. It can be shown (in case M and M' are finite) that $M, w \leftrightarroweq M', w'$ iff for all $\varphi \in \mathcal{L}$, $M, w \models \varphi$ iff $M', w' \models \varphi$.

3.1.3 Characterization of finite models

Finally, we will also use the following proposition.

PROPOSITION 1. *[3][10] Let M be a* finite *epistemic model and $w \in M$. Then there is an epistemic formula $\delta_M(w)$ (involving common knowledge) such that*

1. *$M, w \models \delta_M(w)$*
2. *For every finite epistemic model M' and world $w' \in M'$, if $M', w' \models \delta_M(w)$ then $M, w \leftrightarroweq M', w'$.*

This proposition tells us that a finite epistemic model can be completely characterized (modulo bisimulation) by an epistemic formula. For example, the pointed epistemic model (M, w) in Figure 4 (on page 3) is characterized by the following epistemic formula: $\delta_M(w) := p \wedge (\neg B_Y p \wedge \neg B_Y \neg p) \wedge (\neg B_A p \wedge \neg B_A \neg p) \wedge C_G((\neg B_Y p \wedge \neg B_Y \neg p) \wedge (\neg B_A p \wedge \neg B_A \neg p))$.

This proposition will be very useful to prove that the results of the single agent case transfer to the multi-agent case[2].

3.2 From Possible World to Multi-agent Possible World

3.2.1 The notion of multi-agent possible world

In the AGM framework, one considers a single agent Y. The possible worlds introduced are supposed to represent how the agent Y perceives the surrounding world. Because she is the only agent, these possible worlds deal only with propositional facts about the surrounding world. Now, because 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 introduce a modal structure to our possible worlds. We do so as follows.

[2]Note that van Benthem, in [10], already mentioned that this proposition could be used in belief revision theory

a (single-agent) possible world:

$w : p, \neg q$

a multi-agent possible world:

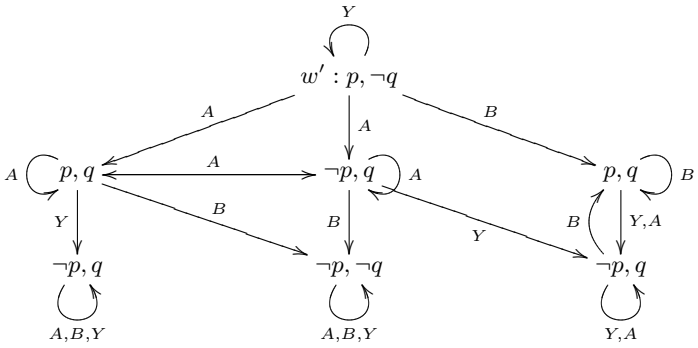

Figure 1: From possible world to multi-agent possible world

DEFINITION 2. *A* multi-agent possible world *(M, w) is a* finite *epistemic model $M = (W, \{R_j; j \in G\}, val)$ such that for all j, R_j is serial, transitive and euclidean, and*

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

Let us have a closer look at the definition. The second condition will be motivated in the next section. The first condition ensures us that in case Y is the only agent then a multi-agent possible world boils down to an interpretation, as in the AGM theory. The first condition also ensures us that in case Y assumes that she is in the multi-agent possible world (M, w) then for her w is the only possible 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). Note that if we remove the constraints on the accessibility relations (seriality, euclidicity and transitivity) the results in this paper are still valid. We prefer to keep them because we find them more intuitive to model the notion of belief construed as conviction. 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 [9] or to Gärdenfors' notion of acceptance [6] or to Voorbraak's notion of rational introspective belief [12].

We see in Figure 1 that a multi-agent possible world is really a generalization of a possible world (or interpretation). In both of them agent Y believes that p is true and q is false. But in the multi-agent possible world we can also express what (agent Y's beliefs about) the other agents' beliefs are. For example, (agent Y believes that) agent A believes that q is true ($B_A q$) or agent B believes that agent Y believes that p is false ($B_B B_Y \neg p$). Inspiring ourselves from the AGM theory we can then define the notion of internal model.

DEFINITION 3. *An* internal model *is a finite and disjoint union of multi-agent possible worlds.*

Note that in the single-agent case, an internal model boils down to a (non-empty) set of interpretations, so represents a belief set. Intuitively, an internal model is the formal model that agent Y has "in her head" and that represents how she perceives the surrounding world. This interpretation differs from Hintikka epistemic models (M, w_a), usually encountered in epistemic logic, which are supposed to represent objectively and from an external point of view how all the agents perceive the actual world w_a.

Example 1. In Figure 2 is depicted an example of internal model. In this internal model, the agent Y does not know wether p is true or not (formally $\neg B_Y p \wedge \neg B_Y \neg p$). Indeed, $M_1, w \models p$ and $M_2, v \models \neg p$. The agent Y also believes that the agent A does not know whether p is true or false (formally $B_Y(\neg B_A p \wedge \neg B_A \neg p)$). Indeed, $M_1, w \models \neg B_A p \wedge \neg B_A \neg p$ and $M_2, v \models \neg B_A p \wedge \neg B_A \neg p$. Finally, the agent Y believes that A believes that she 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 $M_1, w \models B_A(\neg B_Y p \wedge \neg B_Y \neg p)$ and $M_2, v \models B_A(\neg B_Y p \wedge \neg B_Y \neg p)$.

3.2.2 Alternative representation of internal models

DEFINITION 4. *Let* $\{(M_1, w_1), \ldots, (M_n, w_n)\}$ *be an internal model. The* epistemic model associated with $\{(M_1, w_1), \ldots, (M_n, w_n)\}$ *is the* $KD45_G$ *epistemic model* $M = (W, \{R_j; j \in G\}, val)$ *defined as follows.*

- $W := W_1 \cup \ldots \cup W_n$;
- $R_j := R_j^1 \cup \ldots \cup R_j^n$ *for* $j \neq Y$;
- $R_Y := R_Y^1 \cup \ldots \cup R_Y^n \cup \{(w_i, w_k); i, k = 1 \ldots n\}$;
- $val(w) := val_i(w)$ *if* $w \in W_i$.

Example 2. In Figure 2 is represented the internal model $\{(M_1, w), (M_2, v)\}$ and in Figure 4 is represented an epistemic model bisimilar to the epistemic model associated to $\{(M_1, w), (M_2, v)\}$ of Figure 3.

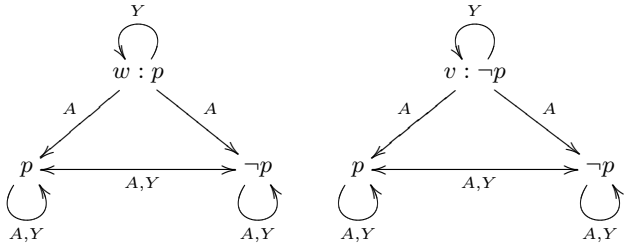

Figure 2: An internal model : multi-agent possible world (M_1, w) (left) and multi-agent possible world (M_2, v) (right)

We can now motivate the second item of Definition 2. Indeed, if this item was not fulfilled then part of the agents j's beliefs about Y's beliefs (for $j \neq Y$) would depend on the other multi-agent possible worlds of the internal model.

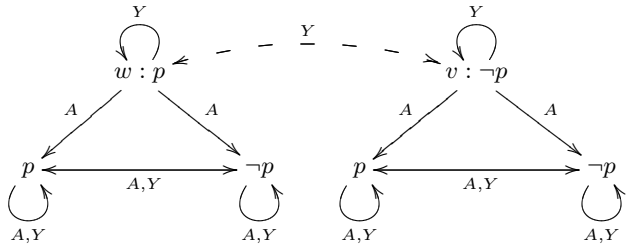

Figure 3: Epistemic model associated to the internal model $\{(M_1, w), (M_2, v)\}$

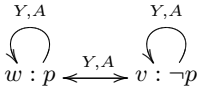

Figure 4: Epistemic model bisimilar to the epistemic model of Figure 3

This aspect of the notion of internal model is revealed when we define the notion of epistemic model associated to an internal model. Condition 2 ensures us that the agents j's beliefs in a multi-agent possible world of a given internal model depend only on the structure of this multi-agent possible world. Condition 2 thus provides a kind of modularity to multi-agent possible worlds that will be useful in the sequel.

For every internal model, the epistemic model associated to this internal model is a $KD45_G$ epistemic model generated by $R_Y(w)$ (for some world w of this epistemic model). The other way round, one can easily show that any $KD45_G$ epistemic model generated from $R_Y(w)$ (for some world w of this epistemic model) can be equivalently represented by an internal model[3]. So we have two equivalent ways to represent the epistemic state of agent Y.

The second type of representation is much closer to usual epistemic models of standard epistemic logic. But we stress that the interpretation of our models are different from the interpretation of epistemic models in standard epistemic logic. Our models are built by agent Y in order to represent for herself the surrounding world, whereas the models of epistemic logic are built by an external modeler and represent truthfully how each agent perceives the actual world.

Besides, the shape of internal models, based on the notion of multi-agent possible world, allows to generalize easily concepts and methods from AGM belief revision theory, as we will now see.

3.3 The Multi-agent Generalization of the AGM Approach

In the multi-agent case like in the single-agent case, it does not make any sense to revise by formulas dealing with what the agent Y believes or considers possible. Indeed, due to the fact that positive and negative introspection are valid in KD45, Y already knows all she believes and all she disbe-

[3]This equivalence could be easily specified formally by stating that for all i and $\varphi \in \mathcal{L}_{\neq Y}$, $M, w_i \models \varphi$ iff $M_i, w_i \models \varphi$, where $\mathcal{L}_{\neq Y}$ is defined in Definition 5.

lieves. So we restrict the epistemic language to a fragment that we call $\mathcal{L}_{\neq Y}$ defined as follows.

DEFINITION 5.

$$\mathcal{L}_{\neq Y} : \varphi ::= \top \mid p \mid B_j\psi \mid \varphi \wedge \varphi \mid \neg\varphi,$$

where p ranges over Φ, ψ ranges over $\mathcal{L}$ and j over $G - \{Y\}$.

We can then apply with some slight modifications the procedure spelled out for the single agent case in Section 2.

First the postulates for multi-agent belief revision are identical to the ones spelled out in Lemma 1 but this time ψ, μ and φ belong to $\mathcal{L}_{\neq Y}$.

Now we define $\mathcal{I}_G$ to be the set of all multi-agent possible worlds modulo bisimulation; more precisely $\mathcal{I}_G$ is made of the smallest multi-agent possible world among each class of bisimilarly indistinguishable multi-agent possible worlds. We define $Mod(\psi)$ by $Mod(\psi) := \{(M, w) \in \mathcal{I}_G;\ M, w \models \psi\}$. Let $\mathcal{M}$ be an internal model. Thanks to Proposition 1 we can easily prove that

Fact (*) there is a formula $form(\mathcal{M}) \in \mathcal{L}_{\neq Y}$ such that $Mod(form(\mathcal{M})) = \mathcal{M}$.

PROOF. Let (M, w) be a multi-agent possible world. Then we set $\delta^*_M(w) := \bigwedge_{p \in val(w)} p \wedge \bigwedge_{p \notin val(w)} \neg p \wedge \bigwedge_{j \in G - \{Y\}} \left(\bigwedge_{v \in R_j(w)} \neg B_j \neg \delta_M(v) \wedge B_j \left(\bigvee_{v \in R_j(w)} \delta_M(v) \right) \right)$. Clearly $\delta^*_M(w) \in \mathcal{L}_{\neq Y}$ and $M, w \models \delta^*_M(w)$. Besides, for all multi-agent possible world $(M', w') \in \mathcal{I}_G$, if $M', w' \models \delta^*_M(w)$ then $M, w \leftrightarrows M', w'$ by applying Proposition 1; so $(M, w) = (M', w')$ by definition of $\mathcal{I}_G$. Let $\mathcal{M} := \{(M_1, w_1), \ldots, (M_n, w_n)\}$. We set $form(\mathcal{M}) := \delta^*_{M_1}(w_1) \vee \ldots \vee \delta^*_{M_n}(w_n)$. Then $form(\mathcal{M}) \in \mathcal{L}_{\neq Y}$ and $Mod(form(\mathcal{M})) = \mathcal{M}$. □

The proof of this fact is made possible because of the modularity of multi-agent possible worlds enforced by condition 2 in our definition of multi-agent possible world. Therefore, this is another motivation for this condition.

We then get the multi-agent generalization of Theorem 1 by replacing interpretations I by multi-agent possible worlds (M, w).

THEOREM 2. *Revision operator $\circ$ on $\mathcal{L}_{\neq Y}$ satisfies conditions $(R1) - (R6)$ iff there exists a faithful assignment that maps each belief base ψ to a total pre-order $\leq_\psi$ defined on $\mathcal{I}_G$ such that $Mod(\psi \circ \mu) = Min(Mod(\mu), \leq_\psi)$.*

PROOF. The proof follows the line of that of Theorem 1. It relies heavily on the fact (*).

The "if" direction is straightforward. For the "only-if" direction, the key is the definition of a faithful assignment for each belief base in terms of $\circ$. For any multi-agent possible world (M, w) and (M', w') ($(M, w) = (M', w')$is permitted), we define a relation $\leq_\psi$ as $(M, w) \leq_\psi (M', w')$ iff either $(M, w) \in Mod(\psi)$

or $(M, w) \in Mod(\psi \circ form(\{(M, w), (M', w')\}))$.

This definition of the assignment is identical to the single agent case. □

REMARK 1. *We have picked only one of the theorems of [8] but in fact all the theorems present in [8] transfer to the multi-agent case. It includes in particular the theorem about $\leq_\psi$ being a partial order instead of a total order.*

In summary, the concept of internal model (more precisely of multi-agent possible world) allows for a straightforward transfer of the AGM framework and results.

4. SOME CONSIDERATIONS SPECIFIC TO OUR MULTI-AGENT APPROACH

In this section we are going to investigate some multi-agent rationality postulates. Indeed, because we add a multi-agent structure to our possible worlds, it is natural to study how (agent Y's beliefs about) the other agents' beliefs evolve during a revision process.

As said in the introduction, the events we study are private announcement made to Y, the other agents not being aware of anything. So, unlike public announcements, the beliefs of the other agents actually do not change and agent Y knows this. Consequently, agent Y's beliefs about the agents who are not concerned by the formula announced to her should not change as well. So, first of all, we need to define formally what are the agents who are concerned by a formula.

4.1 On the kind of information a formula is about

First note that an input may not only concern agents but also the objective state of nature, i.e. propositional facts, that we note $\mathtt{pf}$ and that we will consider as a "Nature" agent. For example, the formula $p \wedge B_j B_i \neg p$ concerns agent j's beliefs but also propositional facts (namely p). Besides, a formula cannot be about Y's beliefs because $\varphi \in \mathcal{L}_{\neq Y}$ by assumption. So what an input is about includes propositional facts but excludes agent Y's beliefs. This leads us to the following definition.

DEFINITION 6. *Let $C_0 := (G \cup \{\mathtt{pf}\}) - \{Y\}$.*
We define by induction the agents who are concerned *by a formula as follows:*

- $C(p) := \mathtt{pf}$; $C(B_j\varphi) := \{j\}$; $C(C_{G_1}\varphi) := G_1$;
- $C(\neg\varphi) := C(\varphi)$; $C(\varphi \wedge \varphi') := C(\varphi) \cup C(\varphi')$.

For example, $C(p \vee (q \wedge B_j B_i r) \wedge B_k r)) = \{\mathtt{pf}, j, k\}$, and $C(B_i p \vee B_j B_k \neg p) = \{i, j\}$.

We then define a language $\mathcal{L}_{C_1}$ whose formulas concern only agents in C_1, and possibly propositional facts if $\mathtt{pf} \in C_1$.

DEFINITION 7. *Let $C_1 \subseteq C_0$. We define the language $\mathcal{L}_{C_1}$ as follows.*

$$\varphi := \top \mid A \mid B_j\psi \mid \varphi \wedge \varphi \mid \neg\varphi,$$

where j ranges over C_1 and ψ over formulas of $\mathcal{L}$. Besides, $A = \Phi$ if $\mathtt{pf} \in C_1$ and $A = \emptyset$ otherwise.

Now we define a notion supposed to tell us whether two pointed and finite epistemic models contain the same information about some agents' beliefs and possibly about propositional facts.

DEFINITION 8. *Let $C_1 \subseteq C_0$. We say that (M, w) and (M', w') are* C_1-bisimilar, *noted $M, w \leftrightarrows_{C_1} M', w'$, iff*

- *if $\mathtt{pf} \in C_1$ then $val(w) = val(w')$ and*

- *for all* $j \in C_1$,

  if $v \in R_j(w)$ *then there is* $v' \in R_j(w')$ *such that* $M, v \leftrightarrows M', v'$,

  if $v' \in R_j(w')$ *then there is* $v \in R_j(w)$ *such that* $M, v \leftrightarrows M', v'$.

PROPOSITION 2. *Let* $C_1 \subseteq C_0$. *Then* $M, w \leftrightarrows_{C_1} M', w'$ *iff for all* $\varphi \in \mathcal{L}_{C_1}$, $M, w \models \varphi$ *iff* $M', w' \models \varphi$.

PROOF. We assume that $\mathtt{pf} \in C_1$, the proof without this assumption is essentially the same.

Assume $M, w \leftrightarrows_{C_1} M', w'$. We are going to prove by induction on $\varphi \in \mathcal{L}_{C_1}$ that $M, w \models \varphi$ iff $M', w' \models \varphi$.

- $\varphi := p$. As $\mathtt{pf} \in C_1$, $val(w) = val(w')$ so $M, w \models p$ iff $M', w' \models p$.
- $\varphi := \varphi_1 \wedge \varphi_2, \varphi := \neg\varphi'$ work by induction hypothesis.
- $\varphi := B_j\varphi', j \in C_1$. Assume $M, w \models B_j\varphi'$ then for all $v \in R_j$, $M, v \models \varphi'$ (*). But for all $v' \in R_j(w')$ there is $v \in R_j(w)$ such that $M, v \leftrightarrows M', v'$.

  So for all $v' \in R_j(w'), M', v' \models \varphi'$ by property of the bisimulation and (*). Finally $M', w' \models B_j\varphi'$, i.e. $M', w' \models \varphi$.

  The other way around we could show that if $M', w' \models B_j\varphi$ then $M, w \models B_j\varphi$.

Assume that for all $\varphi \in \mathcal{L}_{C_1}$, $M, w \models \varphi$ iff $M', w' \models \varphi$ (*)

- clearly $val(w) = val(w')$
- Let $j \in C_1$ and $v \in R_j(w)$.

  Assume for all $v' \in R_j(w')$ it is not the case that $M, v \leftrightarrows M', v'$ (**).

  Then for all $v' \in R_j(w')$ there is $\varphi(v') \in \mathcal{L}$ such that $M, v \models \neg\varphi(v')$ and $M', v' \models \varphi(v')$.

  Let $\varphi(w') := B_j\left(\bigvee_{v' \in R_j(w')} \varphi(v')\right)$; then $\varphi(w') \in \mathcal{L}_{C_1}$.

  Besides $M', w' \models \varphi(w')$ but $M, w \models \neg\varphi(w')$. This is impossible by (*), so (**) is false.

  The other part of the definition of $\leftrightarrows_{C_1}$ is proved similarly.

□

Proposition 2 ensures us that the notion we just defined captures what we wanted. Its proof uses that the models are finite (otherwise the if direction would not hold). We then have a counterpart of Proposition 1.

PROPOSITION 3. *Let* $C_1 \subseteq C_0$, *let* M *be a finite epistemic models and* $w \in M$. *Then there is* $\delta_M^{C_1}(w)$ *such that*

1. $M, w \models \delta_M^{C_1}(w)$;
2. *for every finite epistemic model* M' *and world* $w' \in M'$, *if* $M', w' \models \delta_M^{C_1}$ *then* $M, w \leftrightarrows_{C_1} M', w'$.

PROOF SKETCH. If $\mathtt{pf} \in C_1$, take

$$\delta_M^{C_1} := \bigwedge_{p \in V(w)} p \wedge \bigwedge_{p \notin V(w)} \neg p \wedge \bigwedge_{j \in C_1} \left(\bigwedge_{v \in R_j(w)} \neg B_j \neg \delta_M(v) \wedge B_j\left(\bigvee_{v \in R_j(w)} \delta_M(v)\right)\right)$$

otherwise if $\mathtt{pf} \notin C_1$, take

$$\delta_M^{C_1} := \bigwedge_{j \in C_1} \left(\bigwedge_{v \in R_j(w)} \neg B_j \neg \delta_M(v) \wedge B_j \bigvee_{v \in R_j(w)} \delta_M(v)\right). \quad \square$$

DEFINITION 9. *Let* $\mathcal{M}$ *and* $\mathcal{M}'$ *be two sets of multi-agent possible worlds, we set* $\mathcal{M} \leftrightarrows_{C_1} \mathcal{M}'$ *iff for all* $(M, w) \in \mathcal{M}$ *there is* $(M', w') \in \mathcal{M}'$ *such that* $M, w \leftrightarrows_{C_1} M', w'$, *and for all* $(M', w') \in \mathcal{M}'$ *there is* $(M, w) \in \mathcal{M}$ *such that* $M, w \leftrightarrows_{C_1} M', w'$.

4.2 Some Postulates Specific to our Multi-agent Approach

As we said before, we study private announcement made to Y, the other agents not being aware of anything. So, in particular, Y's beliefs about the beliefs of the agents who are not concerned by the formula should not change. This can be captured by the following postulate:

(R7) Let $\varphi, \varphi' \in \mathcal{L}_{\neq Y}$ such that $C(\varphi) \cap C(\varphi') = \emptyset$.
If $\psi \to \varphi'$ then $(\psi \circ \varphi) \to \varphi'$

This postulate is the multi-agent version of Parikh and Chopra's postulate [5]. The example of the introduction illustrates this postulate: there $\varphi = \neg p$ and $\varphi' = B_j p \wedge B_j C_G p$. Now the semantic counterpart of (*R*7):

PROPOSITION 4. *A revision operator* $\circ$ *satisfies* (*R*7) *iff for all* $\varphi \in \mathcal{L}_{\neq Y}$, *for all* $(M', w') \in Mod(\psi \circ \varphi)$ *there is* $(M, w) \in Mod(\psi)$ *such that* $M, w \leftrightarrows_{C'} M', w'$, *with* $C' := C_0 - C(\varphi)$.

PROOF. The "if" part is straightforward. Let us prove the "only if" part. Let $\varphi \in \mathcal{L}_{\neq Y}$ and let $(M', w') \in Mod(\psi \circ \varphi)$. Assume that for all $(M, w) \in Mod(\psi)$, it is not the case that $M', w' \leftrightarrows_{C'} M, w$. Then for all $(M, w) \in Mod(\psi)$, $M, w \models \neg\delta_{M'}^{C'}(w')$ by proposition 3. So $\psi \to \neg\delta_{M'}^{C'}(w')$. Then $\psi \circ \varphi \to \neg\delta_{M'}^{C'}(w')$ by application of (*R*7). Hence $M', w' \models \neg\delta_{M'}^{C'}(w')$, which is contradictory. □

Let us consider the converse of (*R*7).

(R8) Let $\varphi, \varphi' \in \mathcal{L}_{\neq Y}$ such that $C(\varphi) \cap C(\varphi') = \emptyset$.
If $\psi \wedge \varphi'$ is satisfiable then $(\psi \circ \varphi) \wedge \varphi'$ is satisfiable.

And the semantic counterpart:

PROPOSITION 5. *A revision operator* $\circ$ *satisfies* (*R*8) *iff for all* $\varphi \in \mathcal{L}_{\neq Y}$, *for all* $(M, w) \in Mod(\psi)$ *there is* $(M', w') \in Mod(\psi \circ \varphi)$ *such that* $M, w \leftrightarrows_{C'} M', w'$, *with* $C' := C_0 - C(\varphi)$.

PROOF. Similar to Proposition 4. □

Unlike (*R*7), (*R*8) is not really suitable for revision because all the worlds representing Y's epistemic state "survive" revision process if (*R*8) is fulfilled. This is not the case

in general because new information can discard some previous possibilities. This is however the case for update where we apply the update process to each world independently (see [7] for an in depth analysis). So ($R8$) is more suitable for an update operation.

In fact ($R8$) can be seen as the multi-agent counterpart of the propositional update postulate ($U8$): consider $\psi := B_i p \vee B_j p$ and $\varphi := \neg B_i p$. Then the revised formula is $\psi \circ \varphi = B_j p \wedge \neg B_i p$ according to postulate ($R2$). But according to postulate ($R8$), after the revision $\neg B_j p$ should be satisfiable because $\psi \wedge \neg B_j p$ was satisfiable.

Postulates ($R7$) and ($R8$) together are equivalent to: for all $\varphi, \varphi' \in \mathcal{L}_{\neq Y}$ such that $C(\varphi) \cap C(\varphi') = \emptyset$, $\psi \rightarrow \varphi'$ iff $(\psi \circ \varphi) \rightarrow \varphi'$. Then

PROPOSITION 6. *A revision operator $\circ$ satisfies ($R7$) and ($R8$) iff for all $\varphi \in \mathcal{L}_{\neq Y}$, $Mod(\psi) \leftrightarrows_{C'} Mod(\psi \circ \varphi)$, with $C' := C_0 - C(\varphi)$.*

PROOF. Follows straightforwardly from Proposition 4 and 5. □

5. CONCLUSION

We have proposed a semantics to adequately represent the agent Y's perception of the surrounding world in a multi-agent setting. This semantics generalizes the single agent one of AGM belief revision theory. Then Proposition 1 has enabled us to also generalize easily the results of AGM belief revision theory to the multi-agent case. Finally, we have studied two additional multi-agent postulates that express more explicitly the fact that we study private announcement.

The power of our approach is that it generalizes all the results of AGM belief revision theory to the multi-agent case, and so thanks to the notion of internal model. In fact, if we consider in particular that there are no other agents than Y then our approach boils down to classical AGM belief revision theory.

In the literature of dynamic epistemic logic, there are works that also deal with private multi-agent belief revision ([2],[4] or [11] for example). However, their modeling approach is quite different from ours. The models built in their work are supposed to represent truthfully the situation from an external and objective point of view, as it is usually done in epistemic logic. So they also study private multi-agent belief revision but from an objective and external point of view. On the other hand, our (internal) models are supposed to be built by agent Y in order to represent for herself the surrounding world. And our revision techniques are also intended to be used by agent Y to revise her beliefs. We claim that our framework is more suitable for designing autonomous agents.

Finally, it would be interesting to investigate other multi-agent rationality postulates. Another line of research would be to study multi-agent update as we have started in Section 4.2. Indeed, the results of [8] about propositional update transfer to the multi-agent case as well.

6. ACKNOWLEDGEMENTS

I thank my PhD supervisors Andreas Herzig and Hans van Ditmarsch for useful comments and discussions. I thank Jérôme Lang for comments on an earlier version of this paper. I thank an anonymous reviewer for her/his extensive comments on my submission.

7. REFERENCES

[1] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *J. Symb. Log.*, 50(2):510–530, 1985.

[2] G. Aucher. A combined system for update logic and belief revision. In M. Barley and N. K. Kasabov, editors, *PRIMA 2004*, volume 3371 of *LNCS*, pages 1–17. Springer, 2004. Revised Selected Papers.

[3] P. Balbiani and A. Herzig. Talkin'bout Kripke models. In *Hylo'07*, Dublin, 2007.

[4] A. Baltag and S. Smets. Conditional doxastic models: A qualitative approach to dynamic belief revision. *Electr. Notes Theor. Comput. Sci.*, 165:5–21, 2006.

[5] S. Chopra and R. Parikh. An inconsistency tolerant model for belief representation and belief revision. In *IJCAI*, pages 192–199, 1999.

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

[7] H. Katsuno and A. O. Mendelzon. On the difference between updating a knowledge base and revising it. In *KR*, pages 387–394, 1991.

[8] H. Katsuno and A. O. Mendelzon. Propositional knowledge base revision and minimal change. *Artif. Intell.*, 52(3):263–294, 1992.

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

[10] J. van Benthem. "One is a Lonely Number": logic and communication. In *Logic Colloquium'02*. ASL & A.K. Peters, 2006.

[11] H. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, 147:229–275, 2005.

[12] F. Voorbraak. *As Far as I know. Epistemic Logic and Uncertainty*. PhD thesis, Utrecht University, 1993.