HAL
open science

# La honte : quand émotion et raisonnement sont liés

Carole Adam, Dominique Longin

## To cite this version:

Carole Adam, Dominique Longin. La honte : quand émotion et raisonnement sont liés. Revue des Sciences et Technologies de l'Information - Série RIA : Revue d'Intelligence Artificielle, 2014, Affects, Compagnons Artificiels et Interactions, 28 (1), pp.43-66. hal-03466648

**HAL Id: hal-03466648**

**https://hal.science/hal-03466648**

Submitted on 6 Dec 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open Archive TOULOUSE Archive Ouverte (OATAO)
OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in : http://oatao.univ-toulouse.fr/
Eprints ID : 12811

**To link to this article** : DOI :10.3166/ria.28.43-66
URL : http://dx.doi.org/10.3166/ria.28.43-66

**To cite this version** : Adam, Carole and Longin, Dominique *La honte : quand émotion et raisonnement sont liés*. (2014) Revue d'Intelligence Artificielle, vol. 28 (n° 1). pp. 43-66. ISSN 0992-499X

Any correspondance concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr

# La honte

## Quand émotion et raisonnement sont liés

**Carole Adam** [1], **Dominique Longin** [2]

1. *Université de Grenoble (UJF)*
*Laboratoire d'Informatique de Grenoble, équipe MAGMA*
*Maison J. Kuntzmann, 110 avenue de la Chimie, 38400 Saint-Martin-d'Hères*
*carole.adam@imag.fr*

2. *CNRS, Université de Toulouse (UPS)*
*Institut de Recherche en Informatique de Toulouse, équipe LILaC*
*118 route de Narbonne, 31062 Toulouse cedex 9*
*Dominique.Longin@irit.fr*

---

RÉSUMÉ. *Certaines émotions qualifiées « de base » dans la littérature sont presque du domaine du réflexe. D'autres sont déclenchées* via *des mécanismes de* pattern matching *sur certains états mentaux (épistémiques et motivationnels le plus souvent) en fonction de la congruence ou non de ces états. D'autres enfin, sont issues de mécanismes cognitifs plus complexes (et à ce titre nous les appelons des émotions complexes) tels des raisonnements contrefactuels (pour la culpabilité et le regret par exemple), des jugements normatifs (pour la honte, la fierté,* etc.*), des évaluations probabilistes du monde (comme dans le cas de l'étonnement),* etc.. *Dans ce qui suit, nous étudions et formalisons l'émotion complexe de honte.*

ABSTRACT. *Some emotions, described as « basic » in the literature, are almost reflexes. Other emotions are triggered* via pattern matching *mechanisms operating on specific mental states (most often epistemic and motivational) to determine the (in)congruence of these states. Yet other emotions come from more or less complex cognitive mechanisms (and we thus call them complex emotions) such as counterfactual reasoning (*e.g. *for guilt or regret), normative judgement (*e.g. *shame, pride,* etc.*), probabilistic evaluations of the world (*e.g. *surprise),* etc.. *In the following, we study and formalise the complex emotion of shame.*

MOTS-CLÉS : *émotions, honte, logique modale.*

KEYWORDS: *emotions, shame, modal logic.*

## 1. Introduction

Classer des émotions est toujours en partie subjectif car leur ressenti varie d'un individu à l'autre. Ainsi, Ortony *et al.* (1988) par exemple, ont renoncé à classer les émotions elles-mêmes et préfèrent classer des *catégories* d'émotions. Ce qui détermine les catégories en question est propre à chaque classification. Un critère qui nous paraît intéressant et que nous appelons « critère de complexité cognitive » est d'analyser les émotions selon une dimension relative à la complexité du traitement cognitif qu'elles nécessitent, de la plus simple à la plus complexe. (C'est également le sens que Ortony *et al.* (1988, p. 28) attribuent à « complexe ».)

D'après ce critère, nous pouvons distinguer trois grandes classes. Les émotions dites « de base » sont presque du domaine du réflexe (comme la peur ou une certaine forme de surprise par exemple). D'autres sont déclenchées *via* des mécanismes de *pattern matching* sur certains états mentaux (épistémiques et motivationnels le plus souvent) en fonction de la congruence ou non de ces états (c'est le cas de la joie, la tristesse, la satisfaction, *etc*.). D'autres enfin, sont issues de mécanismes cognitifs plus complexes tels des raisonnements contrefactuels (pour la culpabilité et le regret par exemple), des jugements normatifs (pour la honte, la fierté, *etc*.), des évaluations probabilistes du monde (comme dans le cas de l'étonnement), *etc*. Nous appelons les éléments de cette troisième catégorie les « émotions complexes ».

Notre but n'est pas ici de proposer une typologie des émotions selon ce critère de complexité cognitive mais d'étudier une émotion complexe particulière : la honte. Cette émotion est parfois qualifiée d'« émotion morale » (car elle est en relation avec la transgression d'une norme morale), ou parfois d'*émotion sociale*. Elster (1999, p. 145) souligne combien les normes sociales ont une influence immensément puissante sur le comportement (*an immensely powerful influence on behavior*). En particulier, la honte nous touche dans ce que nous avons de plus intime, de plus personnel, car elle a une influence certaine sur l'image que nous avons de nous-mêmes et sur la manière dont nous pensons être socialement perçus (Tangney, Dearin, 2002). Pour Elster elle constitue *le* support des normes sociales : par exemple, si un agent viole une norme sociale nous pouvons refuser de traiter avec lui, ce qui peut engendrer chez lui de la honte. Et plus il est manifeste que cela nous coûte de refuser de traiter avec lui, plus nous rendons manifeste la gravité de la violation en question, et plus sa honte sera importante (Elster, 1999, p. 146). En d'autres termes, la honte influence notre comportement social. C'est donc une émotion de première importance, mais paradoxalement très peu étudiée en informatique. Or si un avatar pouvait reconnaître d'une manière ou d'une autre qu'un utilisateur a honte, il pourrait agir en conséquence. Par exemple, si un tuteur intelligent détecte que son élève a honte de parler anglais (par exemple, parce qu'il pense le parler mal), il pourrait décider de mettre en place des stratégies afin de le rassurer. Ce mécanisme peut ainsi fonctionner de manière palliative, mais également de manière anticipative pour faire ou ne pas faire une action donnée (qui, éventuellement, provoquerait de la honte chez soi ou chez l'autre).

Il existe de nombreuses théories de l'émotion. Nous nous plaçons dans le cadre de la vision multi-componentielle de Scherer (1984) où l'émotion est vue comme un « épisode émotionnel » ayant une certaine durée (très courte, mais non instantanée) et une certaine dynamique. Les différentes composantes faisant à peu près l'unanimité dans la communauté des psychologues sont les suivantes : *le sentiment* (le ressenti de l'émotion) ; *la réponse psychophysiologique* (accélération du rythme cardiaque, de la température corporelle, *etc*.) ; *l'expression motrice* (du visage, de la voix, des gestes) ; *la tendance à l'action* (à ne pas confondre avec l'action elle-même) ; *l'évaluation cognitive* (*appraisal* en anglais). Dans les théories de l'évaluation cognitive cette dernière composante constitue le déclencheur des quatre autres ; elle représente le processus cognitif d'évaluation d'un certain événement qui déclenche une réponse émotionnelle différenciée, c'est-à-dire qui détermine si c'est une émotion qui est déclenchée plutôt qu'une autre. L'état mental nécessaire et suffisant à cette évaluation cognitive des émotions est ce que nous appelons par la suite la *structure cognitive de l'émotion*. Elle est au cœur de la naissance et de la nature de l'émotion produite. Une conséquence directe de cette origine « évaluative » de l'émotion est que celle-ci est toujours *à propos de* quelque chose (l'objet de l'évaluation). On pourrait dire que la structure cognitive de l'émotion est un état mental qui, au même titre que la croyance, le désir, l'intention, *etc*. renvoie à un état ou un objet du monde (voir (J. R. Searle, 1983) par exemple pour plus de détails à ce sujet).

Notre but dans cet article est de caractériser la structure cognitive de la honte au travers d'un cadre logique afin de rendre compte de ses propriétés. Afin de ne pas rendre notre étude trop complexe, nous n'étudions pas les aspects liés à son intensité (voir à ce sujet (Lorini, 2011) par exemple). Dans ce qui suit, nous analysons l'émotion de honte (section 2) puis nous présentons notre cadre formel (section 3) ; nous utilisons ensuite ce cadre pour caractériser formellement la honte ainsi que quelques scénarios impliquant un raisonnement sur cette émotion (section 4). Finalement nous évoquons informellement des ressemblances et différences entre la honte et une émotion souvent confondue avec elle, la culpabilité (section 5).

## 2. La honte

La honte a été largement étudiée en psychologie (Tangney *et al.*, 1996 ; Tangney, 1999 ; Tangney, Dearin, 2002 ; Lazarus, 1991 ; Ortony *et al.*, 1988). C'est une émotion perçue comme négative à laquelle nous sommes particulièrement sensibles car, comme le note Lewis (1971), lorsqu'un individu éprouve de la honte il se focalise sur sa personne dans sa globalité, l'atteinte à son image, à sa face. Elster (1999, pp. 152–153) dit que dans le cas de la honte on se voit soi-même comme une personne mauvaise (alors que dans le cas de la culpabilité on se voit comme ayant fait quelque chose de mal). Néanmoins la honte joue un rôle social important : elle a la fonction de médiateur cognitif du comportement social d'un individu, et le sentiment désagréable qu'elle nous fait éprouver nous pousse à supprimer ou à remédier à de possibles dysfonction-

nements dans nos relations avec les autres [1]. Par exemple, un jeune dont les vêtements (achetés par ses parents) ne correspondent pas aux codes vestimentaires du groupe auquel il s'identifie pourrait en éprouver de la honte et chercher à atténuer cet aspect de sa personne (en achetant d'autres vêtements, ou en enlevant les plus visibles). Par là même, il rétablit une certaine normalité dans les relations sociales qu'il entretient avec le groupe en question, en répondant mieux à ses codes [2]. Lazarus souligne que bien qu'elle puisse être vue comme survenant de manière privée et sans personne autour, la honte implique toujours d'autres personnes (Lazarus, 1991, p. 241).

La honte est principalement liée au fait qu'on croit avoir violé une valeur morale internalisée, c'est-à-dire une « valeur morale importante » (Ortony *et al.*, 1988, p. 142–143) qu'on se sent moralement engagé à respecter [3]. Pour Lazarus (1991, p. 240 & 242) la honte implique des pensées ou des actions qui violent une proscription sociale internalisée et où le blâme est pour soi-même (voir aussi Ortony *et al.* (1988, pp. 136–144)). Ainsi, c'est une émotion sociale du fait de la nature sociale des normes violées.

Un autre aspect important est qu'on peut ressentir de la honte face à soi et/ou face à quelqu'un d'autre. L'expression « face à » désigne chez Castelfranchi et Poggi (1990) la personne physiquement présente ou non et face à laquelle nous éprouvons une émotion donnée. (Cela s'applique aussi à un groupe de personnes.) Pour ces auteurs, on ne peut avoir honte face à quelqu'un que si on croit que cette personne connaît la situation honteuse. On peut illustrer ce point de vue en reprenant l'exemple d'Elster (1999, p. 151) à propos de Mathilde de la Mole qui est honteuse d'être amoureuse du fils d'un charpentier (Julien Sorel). Tant qu'elle n'a parlé à personne de son secret, elle n'éprouve de la honte [4] que face à elle-même. Ce n'est qu'à partir du moment où elle pense (à tort ou à raison) que d'autres personnes sont au courant de ses sentiments qu'elle éprouve de la honte également face à eux.

Enfin, dans le cas où l'on ressent de la honte face à quelqu'un d'autre, est-il nécessaire de croire que cette personne est au courant de la situation honteuse ? C'est l'avis de Lazarus (1991, p. 241) pour qui il est seulement nécessaire d'imaginer comment

---

1. « Shame has *the function of cognitive mediator of the individual's social behaviour. (...) Through the unpleasant feelings they inflict they lead one to avoid or remediate possible misfunctioning in one's relationships with other people.* » (Castelfranchi, Poggi, 1990, p. 230).

2. Il est important de noter que, bien qu'il agisse en réponse à la honte qu'il éprouve, cela ne signifie pas pour autant qu'il a fait quelque chose avant cela (et dont il se sentirait responsable) qui a provoqué sa honte. Ainsi dans cet exemple, il n'est pas responsable des vêtements que ses parents lui ont achetés et que, peut-être, ils l'ont forcé à mettre.

3. Typiquement, une valeur (ou une norme) non internalisée (et dont l'agent a connaissance) correspond à une valeur dans laquelle l'agent ne se reconnaît pas, qu'il ne considère pas comme importante à respecter. Par exemple, si on sait qu'il est interdit de télécharger de la musique sur internet mais qu'on en télécharge quand même régulièrement, cela signifie qu'on n'a pas internalisé cette norme, ce n'est pas une de nos valeurs morales. Néanmoins, il peut arriver qu'on viole une norme internalisée, mais lorsque cela arrive, on considère cela comme grave (moralement parlant). (Voir (Conte, Castelfranchi, 1995) pour une théorie générale de l'internalisation des normes.)

4. Du fait qu'elle viole une norme sociale importante à ses yeux, à savoir qu'une femme de la noblesse ne doit pas tomber amoureuse d'une personne de rang social inférieur.

certaines personnes réagiraient *si* elles savaient ce qu'on a fait ou non pour éprouver de la honte. Mais Castelfranchi et Poggi ne sont pas de cet avis et selon eux, on peut se projeter dans l'avenir et *imaginer* la honte qu'on ressentirait si notre entourage était au courant d'une chose dont on a honte, mais dans ce cas la honte face à notre entourage n'est pas réellement ressentie, elle est juste imaginée, fantasmée. Elster (1999, p. 152) impose une condition plus forte en parlant de la « présence des autres » mais il semble que cette condition ne soit pas confirmée par des expériences en psychologie (voir (Tangney, Dearin, 2002, p. 14) par exemple, qui soulignent qu'un nombre substantiel des personnes interrogées a rapporté des expériences de la honte s'étant produites alors qu'elles étaient seules).

Bien que nous n'abordions pas par la suite les tendances à l'action relatives à la honte, il est intéressant de noter que cette émotion conduit à cacher ce qu'on a fait, à minimiser les faits ou leurs conséquences négatives, ou le rôle qu'on y a joué. D'une manière générale, la honte pousse à diminuer toute exposition publique (Lazarus, 1991 ; Elster, 1999 ; Tangney, Dearin, 2002).

La formalisation proposée suit l'analyse de Castelfranchi et Poggi (1990, p. 233) pour qui le fait qu'un agent $i$ ressente de la honte à propos d'un fait $F$ devant un agent $j$ et relativement à un certain critère $C$ d'évaluation de $i$, requiert que les quatre conditions suivantes soient crues par $i$ (nous les mettons en parallèle avec l'exemple tiré de Castelfranchi et Poggi (1990) d'un médecin honteux face à son patient de ne pas connaître un nouveau médicament donné, cette méconnaissance faisant de lui un mauvais médecin) : (1) $F$ est vrai (par exemple, le médecin pense qu'il ne connaît pas un nouveau médicament donné) ; (2) le fait que $F$ soit vrai entraîne une évaluation négative de l'agent $i$ à propos d'un certain critère $C$ (par exemple, le fait que le médecin ne connaisse pas ce nouveau médicament entraîne ou compte pour une évaluation négative du fait qu'il soit un bon médecin) et c'est une croyance commune entre $i$ et $j$ ; (3) le fait que le critère $C$ devrait être vrai est une valeur partagée par $i$ et $j$ (par exemple, le médecin et son patient partagent l'idée selon laquelle idéalement il faut être un bon médecin) et cette valeur partagée est une croyance commune à $i$ et $j$ ; (4) enfin, l'agent $i$ a pour but d'être estimé par $j$ par rapport au critère $C$ (par exemple, le médecin souhaite que son patient l'estime en tant que bon médecin).

On retrouve également ce point chez Lazarus (1991, p. 241) pour qui la honte requiert une personne potentiellement critique (sur l'état négatif dont on a honte) et dont l'approbation est importante pour nous. De plus, l'agent $i$ peut être honteux face à l'agent $j$ même si lui-même ne croit pas que les propriétés (1) et (2) sont vraies (par exemple le médecin peut penser qu'il connaît ce nouveau médicament que son patient pense qu'il ne connaît pas, ou bien il peut penser que ne pas connaître un nouveau médicament ne fait pas de lui un mauvais médecin) ; cependant il est nécessaire qu'il croie (3) afin qu'il se sente concerné par la violation de la valeur morale (par exemple si le médecin ne partage pas l'avis de son patient à propos du fait qu'il est important d'être un bon médecin, il n'a aucune raison d'éprouver de la honte face à ce patient).

## 3. Cadre formel

### *3.1. Langage de base et attitudes mentales*

Soit $AGT$ l'ensemble fini des agents et $2^{AGT*} = 2^{AGT} \setminus \emptyset$. Soit $ATM$ l'ensemble des formules atomiques et $ATM_i \subseteq ATM$ pour tout $i \in AGT$ l'ensemble fini de celles représentant des propriétés de l'agent $i$. Le langage $\mathcal{L}_{\mathcal{SL}}$ de la logique de la honte $\mathcal{SL}$ est défini par la BNF suivante :

$$\varphi :: p \mid p_i \mid \neg\varphi \mid \varphi \vee \varphi \mid MBel_G\, \varphi \mid Goal_i\, \varphi \mid SValue_i\, \varphi$$

où $p \in ATM$, $p_i \in ATM_i$, $i \in AGT$ et $G \in 2^{AGT*}$. Les autres connecteurs classiques ($\top$, $\bot$, $\wedge$, $\rightarrow$ et $\leftrightarrow$) sont définis de manière usuelle. $p_i$ signifie que $p$ est une propriété de l'agent $i$.

$MBel_G\, \varphi$ se lit « le fait que $\varphi$ soit vrai est une croyance mutuelle du groupe d'agents $G$ » (c'est-à-dire que tous les agents du groupe $G$ croient $\varphi$, croient que les autres agents croient aussi $\varphi$, qu'ils croient que les autres croient qu'ils croient $\varphi$, *etc*. à l'infini (Fagin *et al.*, 1995)). Il s'agit ici de la croyance à la Hintikka (1962), correspondant à un savoir subjectif : autrement dit, croire $\varphi$ signifie croire que $\varphi$ est vrai dans le monde réel (*i.e*. il est vrai dans tout monde considéré comme une alternative possible au monde actuel).

$Goal_i\, \varphi$ se lit : « l'agent $i$ a pour but que $\varphi$ ». Il s'agit des buts de (Conte, Castelfranchi, 1995) pouvant provenir de désirs (qui sont intrinsèquement endogènes à une personne), de normes internalisées, ou de buts exogènes qui s'imposent à lui (voir (J. Searle, 2001) pour plus de détails). Ainsi, la satisfaction d'un but ne correspond pas nécessairement à un état positif pour l'agent mais plus généralement à un état « moins mauvais » que celui atteint en ne satisfaisant pas ce but. De plus, les buts ne sont pas nécessairement réalistes (au sens où un agent peut avoir un certain but sans pour autant croire que celui-ci pourra être un jour atteint). Par contre, contrairement aux désirs, ils ne peuvent pas être contradictoires entre eux.

$SValue_i\, \varphi$ se lit : « $\varphi$ est une valeur morale de l'agent $i$ particulièrement importante pour lui ». Cela signifie que $\varphi$ est une valeur morale internalisée par $i$, que $i$ s'ordonne de faire en sorte de respecter (Castaneda, 1975). En ce sens, $i$ est moralement responsable de la réalisation de $\varphi$ (ce qui ne signifie pas qu'il va systématiquement avoir pour but de réaliser $\varphi$, mais s'il a le but contraire cela amènera en lui un conflit moral important). Le fait que cela représente une valeur morale particulièrement importante pour $i$ est cohérent avec le type de normes internalisées décrites par Ortony *et al.* (1988, p. 142–143) ou Castelfranchi et Poggi (1990). Ce type de normes est susceptible de faire perdre la face à l'agent en cas de violation.

Il est important de noter que comme les buts peuvent provenir à la fois de désirs (qu'on souhaite voir réalisés ici et maintenant), et de valeurs morales (qu'on souhaite voir respectées ici et maintenant), il peut arriver de façon contingente qu'une valeur morale implique un but. Cependant ce n'est pas nécessaire, car nous souhaitons laisser aux agents la possibilité d'avoir des buts contraires à leurs valeurs morales (ce qui

peut toujours arriver, par exemple pour des impératifs vitaux). Nous supposons qu'un processus de filtrage (non décrit ici) est à l'origine de la formation des buts. (Bien sûr, les croyances de l'agent au moment de ce filtrage interviennent aussi.)

La figure 1 définit un certain nombre d'abréviations. $beLiked_i(j) \in ATM_i$ signifie que l'agent $i$ a la propriété d'être évalué positivement aux yeux de l'agent $j$. En particulier, $beLiked_i(i)$ signifie que l'agent $i$ s'auto-évalue positivement. Ainsi, (Déf$_{bLG}$) signifie que l'agent $i$ a la propriété d'être évalué positivement aux yeux de chaque agent du groupe $G$. (Déf$_{pG}$) signifie que la propriété $p$ est partagée par tous les agents du groupe $G$. (Déf$_{p\emptyset}$) signifie qu'aucun agent de $AGT$ n'a la propriété $p$. (Déf$_{Bel_G}$) se lit: $\varphi$ est une croyance partagée par tous les agents du groupe $G$. (Déf$_{Bel_i}$) se lit: l'agent $i$ croit que $\varphi$ est vrai. (Déf$_{Goal_G}$) signifie que $\varphi$ est un but partagé par tous les agents du groupe $G$. (Déf$_{SValue_G}$) signifie que $\varphi$ est une valeur morale partagée par tous les agents de $G$ et particulièrement importante pour eux.

$$beLiked_i(G) \stackrel{déf}{=} \bigwedge_{j \in G} beLiked_i(j) \qquad \text{(Déf}_{bLG}\text{)}$$

$$p_G \stackrel{déf}{=} \bigwedge_{i \in G} p_i \qquad \text{(Déf}_{pG}\text{)}$$

$$p_\emptyset \stackrel{déf}{=} \bigwedge_{i \in AGT} \neg p_i \qquad \text{(Déf}_{p\emptyset}\text{)}$$

$$Bel_G\, \varphi \stackrel{déf}{=} \bigwedge_{i \in G} MBel_{\{i\}}\, \varphi \qquad \text{(Déf}_{Bel_G}\text{)}$$

$$Bel_i\, \varphi \stackrel{déf}{=} MBel_{\{i\}}\, \varphi \stackrel{déf}{=} Bel_{\{i\}}\, \varphi \qquad \text{(Déf}_{Bel_i}\text{)}$$

$$Goal_G\, \varphi \stackrel{déf}{=} \bigwedge_{i \in G} Goal_i\, \varphi \qquad \text{(Déf}_{Goal_G}\text{)}$$

$$SValue_G\, \varphi \stackrel{déf}{=} \bigwedge_{i \in G} SValue_i\, \varphi \qquad \text{(Déf}_{SValue_G}\text{)}$$

*Figure 1. Abréviations du langage où $i, j \in AGT$, $G \in 2^{AGT*}$*

### *3.2. Sémantique*

Les $\mathcal{SL}$-frames sont des tuples $F = \langle W, \mathcal{B}, \mathcal{G}, \mathcal{I} \rangle$ où :

– $W$ est un ensemble non vide de mondes possibles ou d'états ;

– $\mathcal{B} : AGT \longrightarrow W \times W$ fait correspondre chaque agent $i$ à une relation transitive et euclidienne $\mathcal{B}_i \subseteq W \times W$ entre mondes possibles ;

– $\mathcal{G} : AGT \longrightarrow W \times W$ fait correspondre chaque agent $i$ à une relation sérielle $\mathcal{G}_i \subseteq W \times W$ entre mondes possibles ;

– $\mathcal{I} : AGT \longrightarrow W \times W$ fait correspondre chaque agent $i$ à une relation sérielle $\mathcal{I}_i \subseteq W \times W$ entre mondes possibles.

Dans ce qui suit, nous notons $\mathcal{R}(w) = \{w' \in W : (w, w') \in \mathcal{R}\}$.

$\mathcal{B}_i(w)$ est l'état de croyance de l'agent $i$ dans le monde $w$ : c'est l'ensemble des mondes qu'il considère depuis le monde $w$ comme des alternatives possibles à ce monde $w$. Le fait que chaque relation $\mathcal{B}_i$ soit transitive et euclidienne signifie qu'un agent est conscient de ses croyances et de ses non croyances (voir la contrainte **(SC1)** de la figure 2)[5].

$\mathcal{G}_i(w)$ est l'état motivationnel de l'agent $i$ dans le monde $w$ : c'est l'ensemble des mondes qu'il préfère voir atteints depuis le monde $w$. Ces mondes peuvent ou non provenir de mondes idéaux ou de mondes désirés. Le fait que chaque relation $\mathcal{G}_i$ soit sérielle signifie que les buts de chaque agent ne sont pas contradictoires (il ne peut avoir simultanément $\varphi$ et $\neg\varphi$ comme buts **(SC2)**).

$\mathcal{I}_i(w)$ est l'état des idéaux particulièrement importants de l'agent $i$ dans le monde $w$ : c'est l'ensemble des mondes que l'agent $i$ considère comme moralement super idéaux (c'est-à-dire dont les valeurs morale portées sont parmi les plus importantes à respecter du point de vue de l'agent) depuis le monde $w$. Le fait que chaque relation $\mathcal{I}_i$ soit sérielle signifie que l'agent n'a pas d'idéaux contradictoires **(SC3)**.

Nous imposons également que chaque agent soit conscient de ses buts **(SC4)** et de ses valeurs morales **(SC5)**. Sémantiquement, cela se traduit par le fait que tout monde associé à un but ou à une valeur morale depuis un monde épistémique est également accessible directement depuis le monde courant. Autrement dit, les ensembles des mondes idéaux et motivationnels accessibles depuis les mondes épistémiques et depuis le monde courant sont les mêmes. Cela traduit le fait que si un agent a un but ou une valeur morale, alors il sait qu'il l'a (complétude de la croyance par rapport à ses buts et valeurs morales réelles) et inversement, s'il croit qu'il a un but ou une valeur morale alors il l'a réellement (adéquation de ses croyances par rapport à ses buts et valeurs morales réelles).

Les $\mathcal{SL}$-modèles sont des tuples : $M = \langle F, V \rangle$ où $F$ est une $\mathcal{SL}$-frame et $V : ATM \longrightarrow 2^W$ est une fonction de valuation. Pour toute formule $\varphi$, tout modèle $M$ et tout monde $w$ de ce modèle, nous écrivons $M, w \Vdash \varphi$ qui se lit « $\varphi$ est vrai dans le monde $w$ du modèle $M$ ». On note $M, w \nVdash \varphi$ le fait que $M, w \Vdash \neg\varphi$. Les conditions de vérité sont donc les suivantes :

– $M, w \Vdash p$ ssi $w \in V(p)$;

5. Traditionnellement, cette relation est aussi sérielle ce qui signifie que $\mathcal{B}_i(w)$ ne peut pas se réduire à l'ensemble vide. Autrement dit, cela impose que si l'agent $i$ croit quelque chose alors il existe nécessairement un monde envisagé par lui et accessible depuis $w$ par $\mathcal{B}$ où cette chose est vraie. Ici, cette contrainte n'est pas imposée et $\mathcal{B}_i(w)$ peut être vide. Cela signifie d'une part qu'un agent peut avoir des croyances contradictoires sans que la logique soit inconsistante, d'autre part cela est techniquement imposé par la sémantique des annonces publiques qui, en supprimant des mondes, peuvent ne laisser aucun monde accessible depuis le monde actuel (ce qui serait impossible avec une relation sérielle).

**(SC1)**. si $w' \in \mathcal{B}_i(w)$ alors $\mathcal{B}_i(w) = \mathcal{B}_i(w')$

**(SC2)**. $\mathcal{G}_i(w) \neq \emptyset$

**(SC3)**. $\mathcal{I}_i(w) \neq \emptyset$

**(SC4)**. si $w' \in \mathcal{B}_i(w)$ alors $\mathcal{G}_i(w) = \mathcal{G}_i(w')$

**(SC5)**. si $w' \in \mathcal{B}_i(w)$ alors $\mathcal{I}_i(w) = \mathcal{I}_i(w')$

*Figure 2. Contraintes sémantiques où $w \in W$, $i \in AGT$*

– $M, w \Vdash \neg\varphi$ ssi ce n'est pas le cas que $M, w \Vdash \varphi$;

– $M, w \Vdash \varphi \wedge \psi$ ssi $M, w \Vdash \varphi$ et $M, w \Vdash \psi$;

– $M, w \Vdash MBel_G\, \varphi$ pour tout $G \in 2^{AGT*}$ ssi $M, w' \Vdash \varphi$ pour tout $w' \in (\bigcup_{i \in G} \mathcal{B}_i)^+(w)$ où $(\bigcup_{i \in G} \mathcal{B}_i)^+$ est la fermeture transitive de l'union des relations d'accessibilité épistémique de $G$ ;

– $M, w \Vdash Goal_i\, \varphi$ ssi $M, w' \Vdash \varphi$ pour tout $w' \in \mathcal{G}_i(w)$;

– $M, w \Vdash SValue_i\, \varphi$ ssi $M, w' \Vdash \varphi$ pour tout $w' \in \mathcal{I}_i(w)$.

Une formule $\varphi$ est vraie dans un $\mathcal{SL}$-modèle $M$ si et seulement $M, w \Vdash \varphi$ pour tout monde $w$ de $M$. $\varphi$ est valide si et seulement si $\varphi$ est vraie dans tout $\mathcal{SL}$-modèle (on note alors $\models_{\mathcal{SL}} \varphi$). $\varphi$ est satisfiable si et seulement si sa négation n'est pas valide.

### *3.3. Axiomatique*

Il découle de la sémantique et de nos définitions que les opérateurs de croyance mutuelle $MBel_G$ pour tout $G \in 2^{AGT*}$ sont définis dans une logique de type K4 et ceux de croyance individuelle $Bel_i$ dans une logique de type K45. [6] ($4_{MBel_G}$) signifie que le fait que $\varphi$ soit une croyance commune au sein du groupe $G$ implique que ce fait soit également une croyance commune de ce groupe. ($4_{Bel_i}$) et ($5_{Bel_i}$) signifient respectivement que si l'agent $i$ croit (resp. ne croit pas) $\varphi$ alors il croit qu'il croit (resp. ne croit pas) $\varphi$. On peut montrer la validité des propriétés suivantes (pour tout $G \in 2^{AGT*}$ et tout $i \in G$) :

$$MBel_G\, \varphi \rightarrow Bel_G\, \varphi \tag{MBel1}$$

$$MBel_G\, \varphi \rightarrow MBel_{G'}\, \varphi \quad \text{pour tout} \quad G' \in 2^{G*} \tag{MBel2}$$

$$MBel_G\, \varphi \leftrightarrow \bigwedge_{i \in G} Bel_i\, MBel_G\, \varphi \tag{MBel3}$$

$$\neg Bel_i\, \varphi \rightarrow \neg Bel_i\, MBel_G\, \varphi \tag{MBel4}$$

$$\neg Bel_i\, MBel_G\, \varphi \rightarrow \neg MBel_G\, \varphi \tag{MBel5}$$

6. Nous utilisons les notations de (Chellas, 1980) pour désigner les différentes logiques. La logique K4 est une logique normale vérifiant en plus l'axiome ($4_\bullet$) et la logique K45 est K4 vérifiant en plus l'axiome ($5_\bullet$). Voir Annexe A page 23 pour plus de détails.

Les opérateurs de but et de valeur morale sont définis dans une logique normale de type KD vérifiant respectivement, en plus des propriétés des logiques normales, les axiomes ($D_{Goal_i}$) et ($D_{SValue_i}$) qui signifient respectivement que l'agent $i$ ne peut avoir deux buts contradictoires ni avoir deux valeurs morales contradictoires (c'est-à-dire que si on a $\varphi$ pour but (ou, resp. pour valeur morale) alors on n'a pas sa négation pour but (resp. pour valeur morale)). Du fait des contraintes sémantiques, la logique $\mathcal{SL}$ vérifie les principes suivants :

$$Goal_i\,\varphi \rightarrow Bel_i\,Goal_i\,\varphi \qquad \text{(PIgoal)}$$

$$\neg Goal_i\,\varphi \rightarrow Bel_i\,\neg Goal_i\,\varphi \qquad \text{(NIgoal)}$$

$$SValue_i\,\varphi \rightarrow Bel_i\,SValue_i\,\varphi \qquad \text{(PIsvalue)}$$

$$\neg SValue_i\,\varphi \rightarrow Bel_i\,\neg SValue_i\,\varphi \qquad \text{(NIsvalue)}$$

Ces propriétés signifient respectivement que si un agent a (resp. n'a pas) un certain but ou une certaine valeur morale, alors il croit qu'il l'a (resp. qu'il ne l'a pas).

### *3.4. Extension aux annonces publiques*

On étend le langage de la logique $\mathcal{SL}$ par des opérateurs modaux d'annonce publique (van Ditmarsch *et al.*, 2007) en ajoutant $[\varphi!]\varphi$ à la définition BNF précédente. Nous nous basons sur le cadre défini par Guiraud *et al.* (2009) en l'étendant à des opérateurs de croyance mutuelle et de valeur morale. $[\varphi!]\psi$ se lit « $\psi$ est vrai après l'annonce publique de $\varphi$ ».

La sémantique associée est définie comme mise à jour d'un $\mathcal{SL}$-modèle : la mise à jour de $M = \langle W, \mathcal{B}, \mathcal{G}, \mathcal{I}, V\rangle$ par $\varphi!$ est le modèle $M^{\varphi!} = \langle W^{\varphi!}, \mathcal{B}^{\varphi!}, \mathcal{G}^{\varphi!}, \mathcal{I}^{\varphi!}, V^{\varphi!}\rangle$ tel que :

$$W^{\varphi!} = \{u_b : u \in W\} \cup \{u_c : u \in W\}$$

$$\mathcal{B}^{\varphi!} = \{(u_b, v_b) : v \in \mathcal{B}(u) \text{ et } M, v \Vdash \varphi\} \cup \{(u_c, v_c) : v \in \mathcal{B}(u)\}$$

$$\mathcal{G}^{\varphi!} = \{(u_b, v_c) : v \in \mathcal{G}(u)\} \cup \{(u_c, v_c) : v \in \mathcal{G}(u)\}$$

$$\mathcal{I}^{\varphi!} = \{(u_b, v_c) : v \in \mathcal{I}(u)\} \cup \{(u_c, v_c) : v \in \mathcal{I}(u)\}$$

$$V^{\varphi!}(u_b) = V^{\varphi!}(u_c) = V(u)$$

Intuitivement, les mondes sont dupliqués en deux groupes : celui relatif à la croyance ($u_b$) et celui relatif aux opérateurs de but et de valeur morale ($u_c$). Au niveau des relations d'accessibilité, elles sont intégralement reproduites dans ce dernier groupe alors que dans le premier :

– seuls sont conservés les éléments de la relation épistémique menant à des mondes où la formule annoncée est vraie ;

– les éléments des relations de but et de valeur morale sont dupliqués de manière à ce que le monde de départ soit un monde relatif à la croyance ($u_b$) et celui d'arrivée soit un monde relatif au but ou à une valeur morale ($v_c$).

EXEMPLE 1. — Un exemple est donné figure 3. Pour simplifier, le modèle $M = \langle W, \mathcal{B}, \mathcal{G}, \mathcal{I}, V \rangle$ de départ ne contient que des éléments de $\mathcal{B}_i$ et de $\mathcal{G}_i$ et on suppose que $W = \{u, v_1, v_2, v_3, v_4\}$. Le nouveau modèle issu de l'annonce est $M^{\varphi!}$ dont $W^{\varphi!} = \{u_b, v_{b_1}, v_{b_2}, v_{b_3}, v_{b_4}\} \cup \{u_c, v_{c_1}, v_{c_2}, v_{c_3}, v_{c_4}\}$. On voit que toutes les formules vraies dans $M, u$ sont vraies dans $M^{\varphi!}, u_b$ à l'exception du fait que : $M, u \Vdash \neg Bel_i\, \varphi \wedge \neg Bel_i\, \neg\varphi$ (alors que $M, u \nVdash Bel_i\, \varphi$) et $M^{\varphi!}, u_b \Vdash Bel_i\, \varphi$ (alors que $M^{\varphi!}, u_b \nVdash \neg Bel_i\, \varphi \wedge \neg Bel_i\, \neg\varphi$). Autrement dit, avant la mise à jour l'agent $i$ ne savait pas si $\varphi$ était vrai ou non, et après la mise à jour il croit que $\varphi$ est vrai (il a donc étendu ses croyances). □

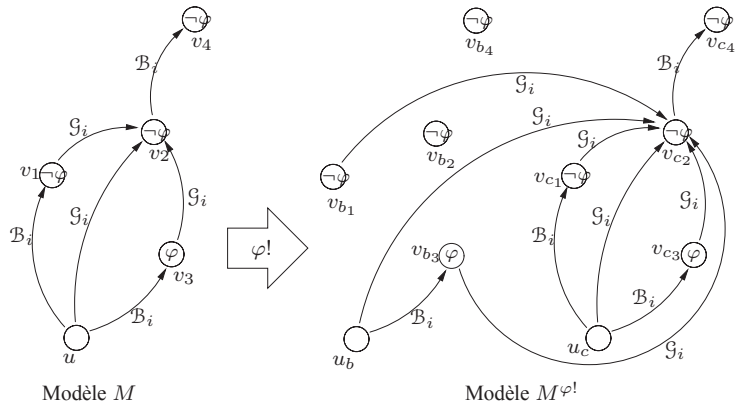


*Figure 3. Exemple où $M, u \Vdash \neg Bel_i\, \varphi \wedge \neg Bel_i\, \neg\varphi$ alors que $M^{\varphi!}, u_b \Vdash Bel_i\, \varphi$*

PROPOSITION 2. — *Pour toute formule $\varphi$, si $M$ est un $\mathcal{SL}$-modèle alors $M^{\varphi!}$ est également un $\mathcal{SL}$-modèle.*

Il faut vérifier la conservation des propriétés sémantiques présentées en FIGURE 2. La preuve est similaire à celle présentée dans (Guiraud *et al.*, 2009). Les démonstrations de **(SC1)**, **(SC2)** et **(SC4)** sont respectivement identiques à celles des contraintes (C.1), (C.2) et (C.3) de (Guiraud *et al.*, 2009). La démonstration de **(SC3)** est similaire à celle de (C.2) et la démonstration de **(SC5)** est similaire à (C.3) : il suffit de remplacer la relation du but par celle d'idéalité (puisque les relations d'idéalité sont contraintes de la même manière que celles des buts).

La condition de vérité associée aux annonces publiques est la suivante :

– $M, u \Vdash [\varphi!]\psi$ ssi $M^{\varphi!}, u_b \Vdash \psi$

Les notions de formule vraie dans un $\mathcal{SL}$-modèle, de validité et de satisfiabilité précédentes sont étendues à la prise en compte des annonces publiques.

On peut montrer que les conditions de vérité associées aux opérateurs rendent les propriétés suivantes valides :

$$\vDash_{\mathcal{SL}} [\varphi!]p \leftrightarrow p \quad \text{où} \quad p \in ATM \tag{RAp}$$

$$\vDash_{\mathcal{SL}} [\varphi!]\neg\psi \leftrightarrow \neg[\varphi!]\psi \tag{RAn}$$

$$\models_{\mathcal{SL}} [\varphi!](\psi_1 \wedge \psi_2) \leftrightarrow [\varphi!]\psi_1 \wedge [\varphi!]\psi_2 \qquad \text{(RAa)}$$

$$\models_{\mathcal{SL}} [\varphi!]Bel_i\,\psi \leftrightarrow Bel_i\,(\varphi \rightarrow [\varphi!]\psi) \qquad \text{(RAb)}$$

$$\models_{\mathcal{SL}} [\varphi!]Goal_i\,\psi \leftrightarrow Goal_i\,\psi \qquad \text{(RAg)}$$

$$\models_{\mathcal{SL}} [\varphi!]SValue_i\,\psi \leftrightarrow SValue_i\,\psi \qquad \text{(RAv)}$$

(RAp) signifie qu'une annonce publique ne change pas les faits, pas plus qu'elle ne change les buts (RAg) ou les valeurs morales (RAv) d'un agent. (RAn) signifie qu'une formule est fausse après une annonce si et seulement si il est faux qu'elle soit vraie après cette annonce. (RAa) signifie que deux faits sont vrais après une annonce si et seulement si chacun des deux est vrai séparément après cette annonce. Enfin, (RAb) signifie qu'une croyance de l'agent $i$ est vraie après une annonce si et seulement si cet agent croit que si le contenu de l'annonce est vrai alors après l'annonce de ce contenu l'objet de cette croyance sera vraie. Les preuves associées aux propriétés (RAp) à (RAg) ont été démontrées dans (Guiraud *et al.*, 2009). La preuve de (RAv) est similaire à celle de (RAg) pour les opérateurs $SValue_i$ .

On peut montrer (cf. Preuve 11) à partir des propriétés précédentes que :

$$\models_{\mathcal{SL}} [\varphi!]\top \qquad (\mathrm{N}_{[\varphi!]})$$

qui signifie que toute annonce préserve les tautologies.

On peut également montrer que les règles d'équivalence suivantes préservent la validité (cf. Preuve 12) :

$$\text{if} \quad \psi \leftrightarrow \psi' \quad \text{then} \quad [\varphi!]\psi \leftrightarrow [\varphi!]\psi' \qquad (\mathrm{RE}_{[\varphi!]})$$

$$\text{if} \quad \varphi \leftrightarrow \varphi' \quad \text{then} \quad [\varphi!]\psi \leftrightarrow [\varphi'!]\psi \qquad (\mathrm{RE}'_{[\varphi!]})$$

Par définition (Chellas, 1980, p. 115) les propriétés (RAa), ($\mathrm{N}_{[\varphi!]}$), ($\mathrm{RE}_{[\varphi!]}$) et (RAn) font que les opérateurs $[\varphi!]$ sont définis dans une logique KD. Les équivalences (RAp) à (RAv) et règles d'inférence ($\mathrm{RE}_{[\varphi!]}$) et ($\mathrm{RE}'_{[\varphi!]}$) ci-dessus sont appelées « des axiomes de réduction » : ils permettent de réduire toute formule $[\varphi!]\varphi$ à une formule ne contenant plus aucun opérateur $[\varphi!]$. Comme cela a été montré par (van Ditmarsch *et al.*, 2007) il n'y a pas d'axiome de réduction pour la croyance commune et l'axiomatique ci-dessus est donc incomplète.

DÉFINITION 3 (formule booléenne et formule positive). — *Pour tout $p \in ATM$, l'ensemble des formules booléennes est tel que:*

$$P ::= p \mid \neg P \mid P \vee P$$

*et l'ensemble des formules positives tel que ($i \in AGT$, $G \in 2^{AGT*}$) :*

$$\varphi^+ ::= P \mid \varphi^+ \vee \varphi^+ \mid \varphi^+ \wedge \varphi^+ \mid MBel_G\,\varphi^+ \mid Goal_i\,\varphi^+ \mid SValue_i\,\varphi^+$$

Par exemple, $Bel_i\,\neg p$, $Goal_i\,MBel_G\,(p \vee \neg q)$ et $p \rightarrow Bel_i\,p$ sont des formules positives, mais pas $Bel_i\,Bel_j\,p \rightarrow Bel_i\,p$ (car cette formule est équivalente à $\neg Bel_i\,Bel_j\,p \vee Bel_i\,p$, et $\neg Bel_i\,Bel_j\,p$ n'est pas une formule positive).

Finalement, nous pouvons démontrer les propriétés suivantes pour tout $i \in AGT$, $G \in 2^{AGT*}$ :

$$[\varphi!]Bel_i\,\varphi \tag{1}$$

$$[\varphi!]MBel_G\,\varphi \tag{2}$$

$$\varphi^+ \rightarrow [\psi!]\varphi^+ \tag{3}$$

$$\neg Bel_i\,\neg P \rightarrow [P!]\neg Bel_i\,\neg P \tag{4}$$

$$Bel_i\,[\varphi!]\psi \rightarrow [\varphi!]Bel_i\,\psi \tag{5}$$

(1) et (2) signifient respectivement qu'après que $\varphi$ ait été annoncé, tous les agents croient (respectivement croient mutuellement) que $\varphi$ est vrai. (3) signifie que toute formule positive reste vraie après n'importe quelle annonce. Voir (van Ditmarsch *et al.*, 2007) pour la démonstration de ces trois propriétés. (4) signifie que si un agent envisage une formule booléenne comme vraie, alors après qu'il ait été annoncé que cette formule est vraie l'agent continue à l'envisager.[7] (5) signifie que si un agent croit qu'après que $\varphi$ ait été annoncé $\psi$ sera vrai, alors après que $\varphi$ ait été annoncé cet agent croira que $\psi$ est vrai. (Cette propriété peut se démontrer facilement à partir de (RAb) et des principes de la logique.)

## 4. Formalisation

DÉFINITION 4. — *Pour tout agent $i \in AGT$, tout groupe d'agents $G \in 2^{AGT*}$ et toute formule $p_i \in ATM_i$ :*

$$\begin{aligned} Shame_i\,(G, \varphi, p_i) \overset{déf}{=} & Bel_i\,MBel_G\,\varphi \wedge \\ & Bel_i\,MBel_G\,(\varphi \rightarrow \neg p_i) \wedge \\ & Bel_i\,MBel_{G\cup\{i\}}\,SValue_{G\cup\{i\}}\,p_i \wedge \\ & Goal_i\,beLiked_i(G) \end{aligned}$$

$Shame_i\,(G, \varphi, p_i)$ se lit : « l'agent a honte devant le groupe $G$ que $\varphi$ soit vrai en relation avec la propriété $p_i$ » et chaque élément de la conjonction correspond respectivement aux propriétés présentées en fin de section 2[8]. En accord avec ces propriétés, l'agent $i$ peut avoir honte :

7. Cela n'est pas une propriété valide pour des formules autres que booléennes. Par exemple, si la formule $P$ en question est $q \wedge \neg Bel_i\,q$ et qu'on suppose que la prémisse du théorème est vraie (*i.e.* $\neg Bel_i\,\neg(q \wedge \neg Bel_i\,q)$ est vrai), alors après l'annonce $(q \wedge \neg Bel_i\,q)!$ il n'est pas vrai que $\neg Bel_i\,\neg(q \wedge \neg Bel_i\,q)$ (car il n'est plus vrai que $\neg Bel_i\,\neg\neg Bel_i\,q$).

8. Selon Castelfranchi et Poggi, $beLiked_i(G)$ devrait dépendre de la propriété $p_i$. Nous avons considéré que cette contrainte est plus générale que ça et que le regard que $j$ porte sur $i$ est important pour $i$ « en général » (et non simplement relativement à $p_i$). (Techniquement, cela ne poserait aucun problème de faire dépendre $beLiked_i(G)$ de $p_i$.)

– face à lui-même seulement (quand $G$ se réduit à $\{i\}$) ;

– face à un groupe $G$ seulement (et non face à lui) auquel il n'appartient pas (quand $i \notin G$) ;

– les deux à la fois (quand $G = G' \cup \{i\}$ avec $G' \neq \emptyset$ et $i \notin G'$).

DÉFINITION 5. — *Pour tout agent $i \in AGT$, tout groupe d'agents $G \in 2^{AGT*}$ et toute formule $p_i \in ATM_i$ :*

$$Shame_i\,(G, \varphi) \stackrel{déf}{=} \bigvee_{p_i \in ATM_i} Shame_i\,(G, \varphi, p_i)$$

$Shame_i\,(G, \varphi)$ se lit : « l'agent a honte devant le groupe $G$ que $\varphi$ soit vrai » et est vraie si et seulement si cet agent a honte devant le groupe $G$ que $\varphi$ soit vrai en relation avec au moins une propriété $p_i$.

Enfin, il est intéressant d'isoler les aspects purement épistémiques de la honte des aspects motivationnels.

DÉFINITION 6. — *Pour tout agent $i \in AGT$, tout groupe d'agents $G \in 2^{AGT*}$ et toute formule $p_i \in ATM_i$ :*

$$\begin{aligned} PShame_i\,(G, \varphi, p_i) \stackrel{déf}{=} & \; Bel_i\, MBel_G\, \varphi \wedge \\ & Bel_i\, MBel_G\, (\varphi \rightarrow \neg p_i) \wedge \\ & Bel_i\, MBel_{G \cup \{i\}}\, SValue_{G \cup \{i\}}\, p_i \end{aligned}$$

$PShame_i\,(G, \varphi, p_i)$ se lit : « l'agent $i$ est susceptible de ressentir de la honte ». Il en ressentira réellement si, de plus, il a ici et maintenant le but de renvoyer une image positive à ceux envers qui il éprouve potentiellement de la honte.

Enfin, le fait que le honte soit définie à partir de croyances de l'agent $i$ signifie que celui-ci peut tout à fait bien se tromper et avoir honte de quelque chose vis-à-vis d'un certain groupe alors même que ce groupe n'a absolument pas les croyances ou les valeurs morales requises.

$$\models_{\mathcal{SL}} Shame_i\,(G, \varphi, p_i) \rightarrow Shame_i\,(\{i\}, \varphi, p_i) \quad \text{ssi} \quad i \in G \tag{SH1}$$

$$\models_{\mathcal{SL}} Shame_i\,(G, \varphi, p_i) \rightarrow Bel_i\, MBel_{G \cup \{i\}}\, SValue_i\, p_i \tag{SH2}$$

$$\models_{\mathcal{SL}} Shame_i\,(G, \varphi, p_i) \rightarrow Shame_i\,(G', \varphi, p_i) \quad \text{pour tout} \quad G' \in 2^{G*} \tag{SH3}$$

$$\models_{\mathcal{SL}} Shame_i\,(G, \varphi, p_i) \leftrightarrow PShame_i\,(G, \varphi, p_i) \wedge Goal_i\, beLiked_i(G) \tag{SH4}$$

$$\models_{\mathcal{SL}} Shame_i\,(G, \varphi, p_i) \rightarrow Bel_i\, Shame_i\,(G, \varphi, p_i) \tag{SH5}$$

$$\models_{\mathcal{SL}} \neg Shame_i\,(G, \varphi, p_i) \rightarrow Bel_i\, \neg Shame_i\,(G, \varphi, p_i) \tag{SH6}$$

(SH1) illustre le fait que si l'agent $i$ a honte face à un groupe $G$ auquel il appartient, alors il a honte face à lui-même. (Voir Preuve 13 en Annexe.) (SH2) ne présuppose pas que $i$ appartienne au groupe $G$ ou non et illustre le fait que même lorsque $i$ a honte face à $G$ auquel il n'appartient pas ($i$ n'a donc pas honte face à lui-même) il doit y avoir croyance commune entre l'agent $i$ et ceux du groupe $G$ que $p_i$ est une valeur morale de $i$. (Preuve triviale ; voir fin de la section 2 pour la justification conceptuelle de cette propriété.) (SH3) rend compte du fait que si un agent est honteux face à un groupe $G$ alors il est honteux face à tout sous-groupe non vide $G'$ de $G$. (La démonstration est immédiate *via* (MBel2), (Déf$_{bLG}$) et les principes de la logiques modales associés à la logique $\mathcal{SL}$.) (SH4) découle directement des définitions. (SH5) et (SH6) illustrent le fait qu'un agent est conscient de ce dont il est honteux et de ce dont il ne l'est pas. Cette propriété découle directement de ($4_{Bel_i}$), ($5_{Bel_i}$) (voir Annexe A), (PIgoal) et (NIgoal).

### *4.1. Exemples*

EXEMPLE 7 (Absence de honte). — Soient $G = \{Tom, Maxime, Kenzo\}$ tel que $G \subseteq AGT$, $desordre \in ATM$ signifiant que la chambre de Tom est en désordre, $cool_{Tom} \in ATM_{Tom}$ signifiant que Tom a la propriété d'être cool, et $cool_{AGT}$ signifiant que tout le monde a la propriété d'être cool. Maxim et Kenzo viennent jouer chez Tom et voient la chambre de ce dernier en désordre ($MBel_G\ desordre$). Aucun d'eux n'estime que le fait d'avoir sa chambre en désordre fait de Tom quelqu'un de « pas cool » ($\bigwedge_{i \in G} \neg Bel_i\,(desordre \rightarrow \neg cool_{Tom})$) et ils croient qu'il est particulièrement important d'être cool ($MBel_G\ SValue_G\ cool_{AGT}$). Enfin, chacun cherche à ce que les autres aient une image positive de lui-même ($\bigwedge_{i \in G} Goal_i \bigwedge_{j \in G \setminus \{i\}} beLiked_i(j)$).

Il est aisé de montrer que Tom n'a pas honte que sa chambre soit en désordre (car il n'estime pas que ce désordre fasse de lui quelqu'un de pas *cool*). Formellement, si on note $KB_7$ l'ensemble de ces faits, cela s'illustre par la validité du principe suivant :$\models_{\mathcal{SL}}$ $KB_7 \rightarrow \neg Shame_{Tom}\,(G, desordre, cool_{Tom})$. □

EXEMPLE 8 (Honte face aux autres). — Dans cet exemple, on remplace les deux premières hypothèses de l'exemple précédent par les trois suivantes (et on conserve les deux dernières) : tout le monde a vu sa chambre rangée ($MBel_G\ rangee$) et Tom croit que Maxim et Kenzo pensent mutuellement que cela fait de lui quelqu'un de pas *cool* ($Bel_{Tom}\ MBel_{\{Maxime, Kenzo\}}\,(rangee \rightarrow \neg cool_{Tom})$), avis qu'il ne partage pas ($\neg Bel_{Tom}\,(rangee \rightarrow \neg cool_{Tom})$). Le reste est inchangé par rapport à l'Exemple 7.

Dans ces conditions, on peut montrer que Tom a honte face à Maxim et Kenzo (mais pas face à lui-même) que sa chambre soit rangée. Formellement, cela s'illustre par la validité des deux principes suivants :

1. $\models_{\mathcal{SL}} KB_8 \rightarrow Shame_{Tom}\,(\{Maxime, Kenzo\}, rangee)$.
2. $\models_{\mathcal{SL}} KB_8 \rightarrow \neg Shame_{Tom}\,(\{Tom\}, rangee)$.

(Preuve 15 p. 25) □

EXEMPLE 9 (Absence de honte par désintérêt de plaire). — Dans cet exemple, Tom croit que son frère Arthur sait que sa chambre est rangée et que pour ce dernier cela fait de lui (Tom) quelqu'un de pas *cool*. Il est évident pour Tom et pour Arthur qu'idéalement on doit être cool. Mais comme il s'agit de son frère, Tom n'a pas en cet instant le but de renvoyer à Arthur une image positive de lui-même. Autrement dit la formule $PShame_i\,(\{Arthur\}, rangee, cool_{Tom}) \wedge \neg Goal_{Tom}\, beLiked_{Tom}(Arthur)$ est vraie). En conséquence, il n'éprouve aucune honte devant son frère du fait que sa chambre soit rangée même si, ce faisant, cela fait de lui quelqu'un de pas *cool*, ce qui est négatif en soi (la formule $Shame_i\,(\{Arthur\}, rangee, cool_{Tom})$ est fausse). □

### *4.2. Dynamique de la honte*

On a montré en introduction de cet article que la honte est une émotion morale. À ce titre, elle peut être le moteur de certains de nos comportements. Nous proposons dans ce qui suit d'illustrer cet aspect formellement. Les réactions que l'on adopte sous le coup de la honte étant dépendantes du contexte, nous fixons ici un cadre qui servira d'exemple illustrant notre propos dans toute la fin de cette section.

EXEMPLE 10 (Honte et évolution des croyances). — On retrouve Tom dont la chambre est en désordre (KB10a) ce dont il est au courant (KB10b). Comme tout adolescent, il souhaite renvoyer une image positive à tous ses amis : Maxim et Kenzo bien sûr, mais également Lila, sa nouvelle petite amie (KB10c). Il sait qu'entre lui et ses deux copains, tous sont d'accord pour considérer qu'on peut avoir sa chambre en désordre et être quand même *cool* (KB10d). Il ne pense pas non plus que ce soit un signe d'immaturité (KB10e) mais il ne connaît pas le point de vue de Lila sur ce point (KB10f).

Soit $AGT = \{Tom, Kenzo, Maxime, Lila\}$ l'ensemble de tous les agents et $ATM = \{desordre, cool_{AGT}, mature_{AGT}\}$ l'ensemble des formules atomiques. La base de connaissance initiale $KB_{10}$ est la suivante :

$$desordre \tag{KB10a}$$

$$Bel_{Tom}\, desordre \tag{KB10b}$$

$$Goal_{Tom}\, beLiked_{Tom}(AGT) \tag{KB10c}$$

$$Bel_{Tom}\, MBel_{\{Tom,Maxime,Kenzo\}}\, \neg(desordre \rightarrow \neg cool_{Tom}) \tag{KB10d}$$

$$Bel_{Tom}\, \neg(desordre \rightarrow \neg mature_{Tom}) \tag{KB10e}$$

$$\begin{aligned} &\neg Bel_{Tom}\, Bel_{Lila}\,(desordre \rightarrow \neg mature_{Tom}) \wedge \\ &\neg Bel_{Tom}\, \neg Bel_{Lila}\,(desordre \rightarrow \neg mature_{Tom}) \end{aligned} \tag{KB10f}$$

$$MBel_{AGT}\, SValue_{AGT}\,(cool_{AGT} \wedge mature_{AGT}) \tag{KB10g}$$

Quand ses amis rentrent dans sa chambre ils voient qu'elle est en désordre (ce que nous assimilons à l'annonce de $desordre$!). Tous les agents le croient donc mu-

tuellement et Tom croit qu'une telle croyance mutuelle existe (i), en particulier Tom continue à le croire (ii). Bien sûr, la chambre est toujours en désordre (ii) et toutes les croyances et buts de Tom sont préservés par l'annonce (iv–viii).

Tous ces faits correspondent aux formules ci-dessous dont nous donnons une indication des règles utilisées pour leur démonstration au sein de notre logique $\mathcal{SL}$.

i. $[desordre!]Bel_{Tom}\ MBel_{AGT}\ desordre$
(par (2), (MBel3) et les $\mathcal{SL}$-principes)

ii. $[desordre!]Bel_{Tom}\ desordre$
(par (KB10b), (3) et les $\mathcal{SL}$-principes)

iii. $[desordre!]desordre$
(par (KB10a), (3) et les $\mathcal{SL}$-principes)

iv. $[desordre!]Goal_{Tom}\ beLiked_{Tom}(AGT)$
(par (KB10c), (3) et les $\mathcal{SL}$-principes)

v. $[desordre!]Bel_{Tom}\ MBel_{\{Tom,Maxime,Kenzo\}}\ \neg(desordre \rightarrow \neg cool_{Tom})$
(par (KB10d), (3) et les $\mathcal{SL}$-principes)

vi. $[desordre!]Bel_{Tom}\ \neg(desordre \rightarrow \neg mature_{Tom})$
(par (KB10e), (3) et les $\mathcal{SL}$-principes)

vii. $[desordre!]\neg Bel_{Tom}\ Bel_{Lila}\ (desordre \rightarrow \neg mature_{Tom})$
(par (KB10f) et les $\mathcal{SL}$-principes)

viii. $[desordre!]Bel_{Tom}\ MBel_{AGT}\ SValue_{AGT}\ (cool_{Tom} \wedge mature_{Tom})$
(par (KB10g), (Déf$_{pG}$), (3), (MBel3) et les $\mathcal{SL}$-principes)

Par définition de la honte, il est immédiat de constater que Tom n'éprouve aucune honte devant qui que ce soit. En revanche, si Lila déclare qu'elle pense que le désordre est un signe d'immaturité (*i.e.* $Bel_{Lila}\ (desordre \rightarrow \neg mature_{Tom})!$ est annoncé), alors Tom va éprouver de la honte face à elle.

En effet, il est aisé de montrer les faits suivants :

ix. $[Bel_{Lila}\ (desordre \rightarrow \neg mature_{Tom})!]Bel_{Tom}\ MBel_{AGT}\ desordre$

x. $[Bel_{Lila}\ (desordre \rightarrow \neg mature_{Tom})!]Bel_{Tom}\ Bel_{Lila}\ (desordre \rightarrow \neg mature_{Tom})$

xi. $[Bel_{Lila}\ (desordre \rightarrow \neg mature_{Tom})!]$
$Bel_{Tom}\ MBel_{AGT}\ SValue_{AGT}\ mature_{Tom}$

xii. $[Bel_{Lila}\ (desordre \rightarrow \neg mature_{Tom})!]Goal_{Tom}\ beLiked_{Tom}(AGT)$

Par définition de la honte, nous avons donc démontré que

$$[Bel_{Lila}\ (desordre \rightarrow \neg mature_{Tom})!]Shame_{Tom}\ (\{Lila\}, desordre, mature_{Tom})$$

En d'autres termes, cela montre que

$$\models_{\mathcal{SL}} KB_{10} \rightarrow [desordre!][Bel_{Lila}\ (desordre \rightarrow \neg mature_{Tom})!]$$
$$Shame_{Tom}\ (\{Lila\}, desordre, mature_{Tom})$$

ce qui signifie que la situation initiale suffit à montrer qu'après que tout le monde ait été mis au courant que la chambre de Tom était désordre, puis que Lila considérait cela comme un manque de maturité, Tom a honte face à Lila que sa chambre ne soit pas rangée car il pense qu'aux yeux de Lila cela remet en cause sa maturité (à lui).

Du fait de (KB10d) on peut également montrer qu'à aucun moment (avant la première annonce, après la première annonce mais avant la seconde, et après la seconde annonce) Tom n'éprouve de la honte face à lui-même, Kenzo et/ou Maxime à propos du désordre de sa chambre (puisque le fait que sa chambre soit en désordre ne fait pas de Tom, à leurs yeux, quelqu'un de pas cool).

$$\models_{\mathcal{SL}} KB_{10} \rightarrow \neg Shame_{Tom}(\{Tom, Kenzo, Maxime\}, desordre, cool_{Tom})$$

$$\models_{\mathcal{SL}} KB_{10} \rightarrow [desordre!] \neg Shame_{Tom}(\{Tom, Kenzo, Maxime\}, desordre, cool_{Tom})$$

$$\models_{\mathcal{SL}} KB_{10} \rightarrow [desordre!][Bel_{Lila}(desordre \rightarrow \neg mature_{Tom})!] \neg Shame_{Tom}(\{Tom, Kenzo, Maxime\}, desordre, cool_{Tom})$$

□

## 5. La culpabilité

La honte et la culpabilité[9] ont été largement étudiées en psychologie (Tangney *et al.*, 1996 ; Tangney, 1999 ; Tangney, Dearin, 2002 ; Lazarus, 1991 ; Ortony *et al.*, 1988) mais ont souvent été confondues ou mal différenciées l'une de l'autre (voir par exemple (Tangney, Dearin, 2002, p. 11–12)). Il est donc intéressant de noter ce qui les rassemble (afin de comprendre pourquoi elles ont pu être confondues) et ce qui les différencie (par ce biais, on souligne les propriétés que la honte n'a pas).

Ces deux émotions ont été confondues car elles partagent certaines propriétés. Par exemple, la culpabilité est, comme la honte, une émotion négative. C'est également une émotion sociale : Baumeister *et al.* (1994, p. 243) avancent que la plupart des instances de culpabilité sont clairement et essentiellement liées à des processus interpersonnels (c'est-à-dire qu'on éprouve de la culpabilité par rapport à quelque chose qu'on a fait à autrui, comme on a honte d'avoir violé une norme importante de la société). Comme la honte, la culpabilité est souvent relative à la violation d'une valeur morale internalisée Lazarus (1991, p. 240 & 242).

Néanmoins, un agent éprouvant de la culpabilité se focalise sur ses (in)actions alors que lorsqu'il éprouve de la honte il se focalise plutôt sur sa personne dans sa globalité, l'atteinte à son image, à sa face (Lewis, 1971). Dans la culpabilité on se voit comme ayant fait quelque chose de mal alors que dans la honte on se voit soi-même comme une personne mauvaise Elster (1999, pp. 152–153). Ortony *et al.*

9. Le terme « culpabilité » renvoie ici à l'émotion et non à une notion juridique.

(1988, p. 142–143) affirment que la culpabilité ne nécessite pas une violation inexcusable d'un standard important (comme c'est le cas dans la honte), mais peut provenir d'une simple responsabilité « technique », où l'individu n'avait pas mesuré les conséquences de son action blâmable. Ortony *et al.* (1988, pp. 136–144) classent la culpabilité dans la catégorie des *self-reproach emotions*, c'est-à-dire des émotions à propos d'actions blâmables que nous avons accomplies. Par ailleurs, la culpabilité est une émotion intra-psychique (*i.e.*, qui se ressent toujours face à soi-même et non face aux autres, même quand elle concerne nos rapports avec eux). Finalement, la culpabilité pousse à vouloir réparer le tort causé par ce que l'on a fait (Elster, 1999, p. 153), ou au moins à contre-balancer les causes du sentiment négatif qu'on a de soi (Lazarus, 1991, p. 243), ou à s'excuser (Baumeister *et al.*, 1994, p. 257) (alors que la honte pousse à réduire l'exposition aux autres, à diminuer l'importance de la situation). Bien sûr, si toutes ces propriétés de la culpabilité ne sont pas des conditions nécessaires (ou caractéristiques) de la honte, cela ne signifie pas qu'elle ne puissent pas être présentes de manière contingentes : ce sont alors des conditions qui ne sont pas incompatibles avec la honte, mais qui ne sont pas nécessaires pour autant.

Une formalisation de la culpabilité et l'étude des liens formels entre culpabilité et honte dépassent le cadre de ce travail, mais une telle formalisation a déjà été faite, y compris dans nos propres travaux (Lorini, Schwarzentruber, 2011 ; Adam *et al.*, 2011 ; Lorini *et al.*, 2014)

## 6. Conclusion

Comme nous l'avons montré, la honte est une émotion relativement complexe (et différente de la culpabilité) en faisant appel à des raisonnements sur les états mentaux des autres agents, la manière dont ils nous considèrent, les torts qu'on a pu leur causer, ou les valeurs morales que nous partageons avec eux et que nous avons violées. C'est une émotion qui trouvera, selon nous, une place naturelle au sein des agents conversationnels en ayant un rôle central dans leurs décisions en situation d'interaction sociale, et plus particulièrement lorsque cette interaction sera entre un agent artificiel et un agent humain.

Plus généralement, de nombreuses études, notamment dans le domaine de la théorie des jeux en économie (voir (Harsanyi, 1982) par exemple) montrent que les individus peuvent être plus au moins sensibles aux sentiments de culpabilité ou de honte (on parle « d'aversion à la honte » ou « d'aversion à la culpabilité »). Nos futurs travaux porteront sur ces aspects, notamment par la formalisation de ces deux émotions au sein d'un même langage formel afin d'étudier les liens entre leurs propriétés formelles, ainsi que les tendances à l'action. Cela devrait permettre l'étude de ces émotions du point de vue de : 1) leur structure cognitive ; 2) leur gestion anticipatoire (on va faire ou ne pas faire cela car on sait qu'après on éprouvera (ou non) de la honte

ou de la culpabilité) [10] ; 3) la gestion du *coping* (quelles actions mettre en œuvre pour diminuer, voire supprimer, l'émotion ressentie). Tous ces aspects seront intégrés, dans le cadre d'un projet sur les organisations sociales, au rôle que peuvent jouer les émotions en général (et la honte en particulier) dans l'intégration d'un membre au sein d'une organisation et à son adaptation aux codes de cette organisation.

*Remerciements*

*Ce travail a été soutenu par les contrats de recherche CECIL N° ANR–08-CORD–005 (`www.irit.fr/CECIL/`) et EmoTES N° ANR–11-EMCO–0004 (`www.irit.fr/EmoTES/`).*

## Bibliographie

Adam C., Gaudou B., Longin D., Lorini E. (2011). Logical modeling of emotions for Ambient Intelligence. In F. Mastrogio, N.-Y. Chong (Eds.), *Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives.* IGI Global.

Baumeister R. F., Stillwell A. M., Heatherton T. F. (1994). Guilt: An Interpersonal Approach. *Psychological Bulletin*, vol. 115, n° 2, p. 243–267.

Castaneda H. N. (1975). *Thinking and doing.* Dordrecht, D. Reidel.

Castelfranchi C., Poggi I. (1990). Blushing as a discourse: Was darwin wrong? In W. R. Crozier (Ed.), *Shyness and embarrassment*, p. 230–251. Cambridge University Press.

Chellas B. F. (1980). *Modal Logic: an Introduction*. Cambridge, Cambridge University Press.

Conte R., Castelfranchi C. (1995). *Cognitive and social action*. London, London University College of London Press.

Elster J. (1999). *Alchemies of the mind: Rationality and the emotions*. Cambridge, Cambridge University Press.

Fagin R., Halpern J. Y., Moses Y., Vardi M. Y. (1995). *Reasonning about knowledge*.

Guiraud N., Herzig A., Lorini E. (2009). *Speech acts as announcements (Dagstuhl Seminar on Information processing, rational belief change and social interaction, Dagstuhl, Germany, 23/08/2009-27/08/2009).* Consulté sur http://drops.dagstuhl.de/opus/volltexte/2009/2293/pdf/09351.HerzigAndreas.Paper.2293.pdf. (Science Publications, Dagstuhl Seminar Proceedings 1862-4405, (en ligne). Also presented at LSIR-2 (workshop at IJCAI 2009).)

Harsanyi J. (1982). Morality and the theory of rational behaviour. In A. K. Sen, B. Williams (Eds.), *Utilitarianism and beyond.* Cambridge, Cambridge University Press.

10. On peut illustrer informellement ce point sur l'Exemple 10 si on suppose que Tom éprouve plutôt de l'aversion à la honte. Comme d'une part il croit que ses amis vont venir dans sa chambre (et découvrir qu'elle est en désordre), et d'autre part il ne sait pas si ce désordre constitue pour Lilas (par qui il souhaite être évalué positivement) un critère d'immaturité ou non, on pourrait associer à Tom un comportement où il anticiperait cette honte en rangeant sa chambre par exemple. Une autre personne, éprouvant une aversion moins forte à la honte, pourrait quant à elle voir les choses de manière plus optimiste et parier sur le fait que Lilas ne verra pas dans le désordre de sa chambre un critère d'immaturité. Enfin, quelqu'un n'ayant qu'une aversion très faible à la honte pourrait être prêt à avoir honte devant Lilas de l'état de sa chambre.

Hintikka J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*. Ithaca, Cornell University Press.

Lazarus R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Lewis H. B. (1971). *Shame and guilt in neurosis*. New-York, International Universities Press.

Lorini E. (2011). A Dynamic Logic of Knowledge, Graded Beliefs and Graded Goals and Its Application to Emotion Modelling. In H. van Ditmarsch, J. Lang, S. Ju (Eds.), *Proceedings of the LORI-III Workshop on Logic, Rationality and Interaction, Guangzhou, P.R.China, 10/10/2011-13/10/2011*, vol. 6953, p. 165–178. Springer-Verlag.

Lorini E., Longin D., Mayor E. (2014). A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation*. ((à paraître))

Lorini E., Schwarzentruber F. (2011). A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, vol. 175, nº 3-4, p. 814–847. (doi:10.1016/j.artint.2010.11.022)

Ortony A., Clore G., Collins A. (1988). *The cognitive structure of emotions*. Cambridge, MA, Cambridge University Press.

Scherer K. (1984). Emotion as a multicomponent process: a model and some cross-cultural data. *Review of personality and social psychology*, vol. 5, p. 37–63.

Searle J. (2001). *Rationality in action*. Cambridge, MIT Press.

Searle J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.

Tangney J. P. (1999). The self-conscious emotions: shame, guilt, embarrassment and pride. In T. Dalgleish, M. Power (Eds.), *Handbook of cognition and emotion.* John Wiley & Sons.

Tangney J. P., Dearin R. L. (2002). *Shame and guilt*. The Guilford Press.

Tangney J. P., Miller R. S., Flicker L., Barlow D. H. (1996). Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology*, vol. 70, nº 6, p. 1256–1269.

van Ditmarsch H. P., van der Hoek W., Kooi B. (2007). *Dynamic epistemic logic*. Kluwer Academic Publishers.

## Annexe A. Principaux schémas d'axiomes de la logique modale

Les principaux schémas axiomes et règles d'inférence de la logique modale utilisé précédemment sont détaillés ci-dessous. Ceux-ci sont donnés pour un opérateur générique $\bullet$ (qui se lit *bullet*).

$$\frac{\varphi}{\bullet\varphi} \qquad (\mathrm{RN}_\bullet)$$

$$\bullet\,(\varphi \wedge \psi) \rightarrow (\bullet\varphi \wedge \bullet\psi) \qquad (\mathrm{M}_\bullet)$$

$$(\bullet\varphi \wedge \bullet\psi) \rightarrow \bullet(\varphi \wedge \psi) \qquad (\mathrm{C}_\bullet)$$

$$\bullet\,(\varphi \rightarrow \psi) \rightarrow (\bullet\varphi \rightarrow \bullet\psi) \qquad (\mathrm{K}_\bullet)$$

$$\bullet\,\varphi \rightarrow \neg\,\bullet\,\neg\varphi \qquad (\mathrm{D}_\bullet)$$

$$\bullet\,\varphi \rightarrow \varphi \qquad (\mathrm{T}_\bullet)$$

$$\bullet\,\varphi \rightarrow \bullet\,\bullet\,\varphi \qquad (4_\bullet)$$

$$\neg\,\bullet\,\varphi \rightarrow \bullet\neg\,\bullet\,\varphi \qquad (5_\bullet)$$

Pour obtenir un principe correspondant à un opérateur donné il suffit de substituer ce dernier à l'opérateur *box* dans les formules précédentes. Ainsi par exemple, ($\mathrm{K}_{Bel_i}$) correspond à ($\mathrm{K}_\bullet$) dans lequel on a substitué $\bullet$ par $Bel_i$ ce qui donne pour ($\mathrm{K}_{Bel_i}$) le schéma d'axiome $Bel_i\,(\varphi \rightarrow \psi) \rightarrow (Bel_i\,\varphi \rightarrow Bel_i\,\psi)$.

On appelle logique modale normale toute logique modale dont l'opérateur $\bullet$ vérifie au moins ($\mathrm{RN}_\bullet$) et ($\mathrm{K}_\bullet$). Une logique de type KD est une logique modale normale dont l'opérateur vérifie en plus ($\mathrm{D}_\bullet$), une logique de type K4 est une logique normale vérifiant en plus l'axiome ($4_\bullet$) *etc*. Une logique communément appelée S5 est une logique de type KT5.

## Annexe B. Preuves

PREUVE 11 (de la propriété $\mathrm{N}_{[\varphi!]}$). — On suppose qu'il existe un modèle $M$ et un monde $w$ de ce modèle tel que $M, w \Vdash \neg[\varphi!]\top$ et on montre qu'on arrive à une inconsistance.

$M, w \Vdash \neg[\varphi!]\top$ ssi $M, w \Vdash [\varphi!]\bot$ (par (RAn) et définition de $\top$) ssi $M, w \Vdash [\varphi!](p \vee \neg p)$ (par définition de $\bot$) ssi $M, w \Vdash [\varphi!]p$ et $M, w \Vdash [\varphi!]\neg p$ (par (RAa) et la logique classique) ssi $M^{\varphi!}, w_b \Vdash p$ et $M^{\varphi!}, w_b \nVdash p$ (par la condition de vérité de $[\varphi!]\psi$ et notre définition de la satisfaction) ssi $p \in V(w_v)$ et $p \notin V(w_v)$, ce qui est contradictoire. ■

PREUVE 12 (des propriétés ($\mathrm{RE}_{[\varphi!]}$) and ($\mathrm{RE}'_{[\varphi!]}$)). — On procède par l'absurde. En supposant que $\vDash_{\mathcal{SL}} \psi \leftrightarrow \psi'$ et en supposant qu'il existe un modèle $M$ et un monde $w$ de ce modèle tels que $M, w \Vdash [\varphi!]\psi \wedge \neg[\varphi!]\psi'$ on montre qu'on arrive à une contradiction. On procède de même pour la réciproque.

Il découle de nos hypothèses qu'il existe $M, w$ tels que $M, w \Vdash [\varphi!]\psi$ et $M, w \Vdash \neg[\varphi!]\psi'$ (par condition de vérité de l'opérateur $\wedge$) ssi il existe $M, w$ tels que $M, w \Vdash [\varphi!]\psi$ et $M, w \Vdash [\varphi!]\neg\psi'$ (par (RAn)) ssi il existe $M^{\varphi!}, w_b$ tels que $M^{\varphi!}, w_b \Vdash \psi$ et $M^{\varphi!}, w_b \Vdash \neg\psi'$ (par condition de vérité de l'opérateur $[\varphi!]$) ssi il existe $M^{\varphi!}, w_b$ tels que $M^{\varphi!}, w_b \Vdash \psi$ et $M^{\varphi!}, w_b \nVdash \psi$ (par hypothèse sur le fait que $\psi \leftrightarrow \psi'$ dans tout $\mathcal{SL}$-modèle et nos définitions de la satisfaction) ssi il existe $M^{\varphi!}, w_b$ tels que $\psi \in V^{\varphi!}(w_b)$ et $\psi \notin V^{\varphi!}(w_b)$ ce qui est contradictoire.

De la même manière on montre que s'il existe $M, w$ tels que $M, w \Vdash [\varphi!]\psi' \wedge \neg[\varphi!]\psi$ on arrive à une contradiction, ce qui suffit à prouver la réciproque de l'implication précédente, donc l'équivalence.

La preuve de ($\mathrm{RE}'_{[\varphi!]}$) est similaire et ne présente pas de difficulté. ■

PREUVE 13 (de la propriété (SH1)). — Par définition, on a que

$$\begin{aligned} Shame_i\,(\{i\}, \varphi, p_i) \overset{déf}{=} & \; Bel_i\, MBel_{\{i\}}\, \varphi \wedge Bel_i\, MBel_{\{i\}}\, (\varphi \to \neg p_i) \wedge \\ & Bel_i\, MBel_{\{i\}}\, SValue_{\{i\}}\, p_i \wedge Goal_i\, beLiked_i(\{i\}) \end{aligned}$$

Dans le cas où $i \in G$, on cherche à montrer que chacun des 4 conjoints de la définition de $Shame_i\,(G, \varphi, p_i)$ implique le conjoint correspondant dans $Shame_i\,(\{i\}, \varphi, p_i)$. Or quand $i \in G$, par (MBel1), (Déf$_{Bel_i}$) et les principes de la logiques modale, on a que $Bel_i\, MBel_G\, \varphi \to Bel_i\, MBel_{\{i\}}\, \varphi$. De la même manière on montre que $Bel_i\, MBel_G\, (\varphi \to \neg p_i) \to Bel_i\, MBel_{\{i\}}\, (\varphi \to \neg p_i)$ et $Bel_i\, MBel_{G\cup\{i\}}\, SValue_{G\cup\{i\}}\, p_i \to Bel_i\, MBel_{\{i\}}\, SValue_{\{i\}}\, p_i$ quand $i \in G$. Par (Déf$_{bLG}$) on a également que $Goal_i\, beLiked_i(G) \to Goal_i\, beLiked_i(\{i\})$.

Dans le cas où $i \notin G$, il faut montrer que

$$\nVdash_{\mathcal{SL}} Shame_i\,(G, \varphi, p_i) \to Shame_i\,(\{i\}, \varphi, p_i)$$

Autrement dit, si $i \notin G$ il faut démontrer qu'il existe un modèle $M$ et un monde $w \in W$ tel que $M, w \Vdash Shame_i\,(G, \varphi, p_i) \wedge \neg Shame_i\,(\{i\}, \varphi, p_i)$, *i.e.* tel que $M, w \Vdash Shame_i\,(G, \varphi, p_i)$ et $M, w \Vdash \neg Shame_i\,(\{i\}, \varphi, p_i)$. Comme $M, w \Vdash \neg Shame_i\,(\{i\}, \varphi, p_i)$ ssi $M, w \Vdash \neg Bel_i\, \varphi$ ou $M, w \Vdash \neg Bel_i\,(\varphi \to \neg p_i)$ ou $M, w \Vdash \neg SValue_i\, p_i$ ou $M, w \Vdash \neg Goal_i\, beLiked_i(\{i\})$. Il suffit de choisir $w$ tel que $M, w \Vdash Shame_i\,(G, \varphi, p_i) \wedge \neg Goal_i\, beLiked_i(\{i\})$ par exemple, ce qui implique en particulier qu'il faille vérifier qu'il existe un modèle $M$ et un monde $w \in W$ tel que $M, w \Vdash Goal_i\, beLiked_i(G) \wedge \neg Goal_i\, beLiked_i(\{i\})$ (les autres propositions de la conjonction définissant $Shame_i\,(G, \varphi, p_i)$ étant exclusivement relatifs à des opérateurs épistémiques, il ne peut y avoir d'inconsistance entre eux et $\neg Goal_i\, beLiked_i(\{i\})$). C'est le cas ssi 1) $M, w \Vdash Goal_i\, beLiked_i(G)$ et 2) $M, w \Vdash \neg Goal_i\, beLiked_i(\{i\})$, ssi 1) pour tout $w' \in W$, si $w' \in \mathcal{G}_i(w)$ alors $M, w' \Vdash beLiked_i(G)$ et 2) il existe $w'' \in W$ tel que $M, w'' \Vdash \neg beLiked_i(\{i\})$. Même dans le cas où $w' = w''$, il est immédiat de constater que si $i \notin G$ alors $M, w' \Vdash beLiked_i(G) \wedge \neg beLiked_i(\{i\})$ est consistant. ■

PREUVE 14 (de l'exemple 7). — Il faut montrer que

$$\models_{\mathcal{SL}} KB_7 \rightarrow \neg Shame_{Tom}\,(G, desordre, cool_{Tom}).$$

$Shame_{Tom}\,(G, desordre, cool_{Tom}) \rightarrow Bel_{Tom}\,MBel_G\,(desordre \rightarrow \neg cool_{Tom})$. Or $KB_7 \rightarrow \bigwedge_{i\in G} \neg Bel_i\,(desordre \rightarrow \neg cool_{Tom})$ et $\bigwedge_{i\in G} \neg Bel_i\,(desordre \rightarrow \neg cool_{Tom}) \rightarrow \neg Bel_{Tom}\,MBel_G\,(desordre \rightarrow \neg cool_{Tom})$ qui par contraposition de la première implication conduit au résultat attendu. ■

PREUVE 15 (de l'exemple 8). — On cherche tout d'abord à montrer que

$$\models_{\mathcal{SL}} KB_8 \rightarrow Shame_{Tom}\,(\{Maxime, Kenzo\}, rangee).$$

D'après Définition 5, $Shame_{Tom}\,(\{Maxime, Kenzo\}, rangee, cool_{Tom}) \rightarrow Shame_{Tom}\,(\{Maxime, Kenzo\}, rangee)$. Il suffit donc de montrer que $KB_8 \rightarrow Shame_{Tom}\,(\{Maxime, Kenzo\}, rangee, cool_{Tom})$ ce qui revient à montrer que $KB_8$ implique les quatre principes suivants puisque ceux-ci impliquent le conséquent de cette implication :

1. $Bel_{Tom}\,MBel_{\{Maxime,Kenzo\}}\,rangee$
2. $Bel_{Tom}\,MBel_{\{Maxime,Kenzo\}}\,(rangee \rightarrow \neg cool_{Tom})$
3. $Bel_{Tom}\,MBel_G\,SValue_G\,cool_{Tom}$
4. $Goal_{Tom}\,beLiked_{Tom}(\{Maxime, Kenzo\})$

1. D'après les propriétés (MBel3) et (MBel2) p. 51 et les principes de la logique modale pour les opérateurs $MBel$ (cf. Annexe A) on obtient $Bel_{Tom}\,MBel_{\{Maxime,Kenzo\}}\,rangee$ à partir de $MBel_G\,rangee$, donc de $KB_8$.

2. Cette proposition peut directement être déduite de $KB_8$.

3. Par (Déf$_{pG}$) et les principes de la logique modale (cf. Annexe A) on a que l'hypothèse $MBel_G\,SValue_G\,cool_{AGT}$ implique $MBel_G\,SValue_G\,cool_{Tom}$, ce qui implique par (MBel3) et les principes de la logique modale $Bel_{Tom}\,MBel_G\,SValue_G\,cool_{Tom}$.

4. Il découle de l'hypothèse $\bigwedge_{i\in G} Goal_i \bigwedge_{j\in G\setminus\{i\}} beLiked_i(j)$ que $Goal_{Tom} \bigwedge_{j\in\{Maxime,Kenzo\}} beLiked_{Tom}(j)$ ce qui entraîne par (Déf$_{bLG}$) et les principes de la logique modale que $Goal_{Tom}\,beLiked_{Tom}(\{Maxime, Kenzo\})$.

Il reste maintenant à démontrer que

$$\models_{\mathcal{SL}} KB_8 \rightarrow \neg Shame_{Tom}\,(\{Tom\}, rangee).$$

De la Définition 4, (Déf$_{Bel_i}$) et les principes de la logique modale il découle que $Shame_{Tom}\,(\{Tom\}, rangee, cool_{Tom}) \rightarrow Bel_{Tom}\,(rangee \rightarrow \neg cool_{Tom})$. Par contraposition, il découle que l'hypothèse $\neg Bel_{Tom}\,(rangee \rightarrow \neg cool_{Tom})$ implique $\neg Shame_{Tom}\,(\{Tom\}, rangee, cool_{Tom})$. Comme $ATM_{Tom}$ se réduit au singleton $\{cool_{Tom}\}$ on a par la Définition 5 ce qu'on souhaite démontrer. ■