PLEIAD, un agent émotionnel pour évaluer la typologie OCC

Carole Adam — Andreas Herzig — Dominique Longin

Université Paul Sabatier
Institut de Recherche en Informatique de Toulouse
UMR 5505 — Équipe LILAC
118 route de Narbonne, F-31500 Toulouse
{adam,herzig,longin}@irit.fr

RÉSUMÉ. *Pour résoudre des problèmes complexes, les agents doivent interagir avec les humains de manière naturelle, en particulier en exprimant des émotions réalistes et pertinentes. Les chercheurs se basent donc sur des théories psychologiques pour développer des modèles informatiques des émotions, et principalement sur la typologie dite OCC (Ortony, Clore et Collins). Dans cet article nous proposons d'évaluer cette théorie en procédant en trois étapes. Nous développons d'abord un formalisme basé sur les logiques BDI permettant d'exprimer vingt émotions de la typologie OCC. Nous l'implémentons ensuite dans un agent BDI nommé PLEIAD. Enfin nous demandons à des utilisateurs humains de juger la pertinence des émotions exprimées par cet agent au cours d'un scénario. Cet article décrit ces trois étapes, ainsi que les conclusions de l'évaluation et les perspectives de continuation.*

ABSTRACT. *More and more often, agents have to communicate with humans to solve complex problems. Now, to be accepted by humans, they must express realist and relevant emotions. Researchers thus build on psychological theories to design computational models of emotions. They mainly use Ortony, Clore and Collins' theory known as the OCC typology. In this paper, we propose to assess this theory, processing in three steps. First, we develop a formalism extending BDI logics allowing to express twenty emotions of the OCC typology. Second, we implement these definitions in a BDI agent named PLEIAD, allowing him to express emotions in response to stimuli. Finally, we ask human users to assess the relevance of the emotions expressed by the agent during a scenario. This paper describes these three steps, the conclusions of the evaluation, and our future prospects.*

MOTS-CLÉS : *émotions, agent intelligent, logique BDI, typologie OCC, psychologie, évaluation.*

KEYWORDS: *emotions, intelligent agent, BDI logic, OCC typology, psychology, evaluation.*

RSTI - RIA – 21/2007. SMA pour des environnements complexes, pages 781 à 811

1. Introduction

Les concepteurs de systèmes multi-agents (SMA) doivent aujourd'hui répondre à de nouveaux défis et adapter leurs systèmes à des environnements de plus en plus complexes. En particulier, ils doivent permettre à leurs SMA de s'ouvrir à des agents humains susceptibles d'interagir avec les agents logiciels. Une bonne coopération entre ces deux types d'agents très différents sera sans doute à la base de la réussite de tâches de plus en plus complexes, ce qui explique l'intérêt pour des langages de communication robustes permettant aux agents artificiels de dialoguer avec des humains (*e.g.* (Sadek, 1991; Sadek, 2000; FIPA, 2002; Berger *et al.*, 2005)). Or, des travaux ont montré que pour se faire accepter par les humains, les agents doivent manifester un comportement social, par le respect des normes sociales en vigueur, mais aussi par l'expression d'émotions ou d'états mentaux (Picard, 1997; Gratch *et al.*, 2005). Ces émotions doivent bien sûr être pertinentes dans le contexte de l'interaction, afin de préserver la crédibilité de l'agent. De nombreux travaux se réfèrent donc à des théories psychologiques pour donner des émotions à leurs agents et garantir la pertinence de leur modèle et des résultats obtenus (Gratch *et al.*, 2005). Parmi ces théories de référence, la plus largement utilisée est la typologie OCC (Ortony *et al.*, 1988), et cela avant tout, selon nous, grâce à sa simplicité qui la rend très accessible aux informaticiens. Néanmoins, force est de constater que du point de vue psychologique rien ne prouve que cette théorie soit plus pertinente, plus réaliste, plus fine, ou meilleure que les autres. Dans un contexte où l'on essaye de modéliser des agents qui soient les plus crédibles et les plus réalistes possibles, il semble pourtant important de se poser cette question, d'autant que les théories psychologiques diffèrent sensiblement, tant du point de vue des émotions recensées que dans les définitions qu'elles donnent de ces émotions. Le but de ce travail est donc d'évaluer la pertinence des émotions que peut exprimer un agent BDI construit à partir de la typologie OCC. Nous entendons par « agent BDI » (pour *belief, desire, intention*) un agent dont l'architecture est fondée sur la modélisation de ses états mentaux (*e.g.* (Wooldridge, 2000)).

Pour cela, nous avons dans un premier temps construit un *modèle théorique* des émotions selon la typologie OCC à l'aide d'une logique BDI (*i.e.* une logique comportant entre autres des opérateurs pour représenter les attitudes mentales des agents). À l'origine, cette logique décrit l'architecture d'agents rationnels pouvant interagir avec leur environnement, et est donc indépendante des émotions modélisées. Selon nous, cela présente, entre autres, l'avantage de pouvoir à terme représenter les émotions de différentes théories psychologiques, donc de pouvoir comparer ces théories. Dans un deuxième temps, nous avons alors implémenté une partie de ce modèle théorique dans un agent logiciel (appelé PLEIAD), capable d'exprimer les émotions qu'il « éprouve » en réaction à des stimuli envoyés par l'utilisateur. L'agent dispose d'un ensemble de vingt émotions parmi les vingt-deux de la typologie OCC. Nous soulignons dès à présent qu'à titre expérimental, PLEIAD intègre une gestion de degrés numériques associés aux attitudes mentales, et des degrés d'intensité associés aux émotions. Enfin, dans un troisième temps, nous avons fait évaluer l'agent PLEIAD

par une quinzaine de personnes[1]. Au cours d'un scénario, elles ont pu noter la pertinence des émotions exprimées par l'agent par rapport aux émotions qu'elles auraient ressenti dans la même situation, ou par rapport aux émotions admises dans cette situation. Le but était triple : (1) tester l'adéquation des émotions générées par l'agent par rapport aux résultats produits par notre modèle théorique ; (2) tester les prédictions de notre modèle théorique par rapport à la typologie OCC ; et (3) tester en dernier ces propriétés de OCC par rapport aux attentes des utilisateurs.

Dans ce qui suit, nous présentons d'abord un état de l'art des théories psychologiques des émotions, illustrant leur diversité (section 2). Nous développons ensuite notre formalisme logique et nos définitions formelles des émotions (section 3). Nous décrivons alors une implémentation de cette théorie dans l'agent logiciel PLEIAD, présentons les résultats de l'évaluation menée grâce à cet agent, et en tirons des conclusions aussi bien sur notre formalisation que sur la typologie OCC (section 4). Pour conclure nous discutons d'autres formalisations existantes des émotions (section 5), puis envisageons les perspectives de nos travaux à plus long terme (section 6).

2. Les fondements de notre modèle

Il existe plusieurs types de modèles des émotions. Les modèles évolutionnistes (Darwin, 1872; Ekman, 1992) sont principalement descriptifs, fournissant des taxonomies d'émotions de base. Les modèles dimensionnels (Russell, 1997) sont appropriés pour décrire la dynamique de l'expression des émotions (Becker *et al.*, 2004) mais n'expliquent pas leur déclenchement. Enfin les théories de l'évaluation cognitive (*appraisal*) insistent sur la détermination cognitive de l'émotion et sa fonction adaptative. Notre logique est basée sur ces dernières. Le terme d'*appraisal* a d'abord été introduit par (Arnold, 1960) pour décrire le déclenchement des émotions, en même temps que la notion de *tendances à l'action* qui décrit leur influence. Par la suite ces deux concepts ont été au centre de nombreuses approches ; nous en citons quelques-unes.

2.1. *Lazarus*

Selon (Lazarus, 1991), les émotions résultent de l'évaluation cognitive de l'interaction entre l'individu et son environnement relativement à ses motivations et ses buts. Lazarus définit deux types d'*appraisal*, réalisés dans n'importe quel ordre : l'*appraisal* primaire s'intéresse à la pertinence et la congruence du stimulus par rapport au bien-être de l'individu ; l'*appraisal* secondaire évalue les ressources disponibles pour faire face à ce stimulus. Comme Arnold, Lazarus considère que les émotions induisent des tendances à l'action qui impliquent des modifications physiologiques pour permettre l'adaptation de l'individu à son environnement. Sa théorie est utilisée par exemple dans l'agent EMA (Gratch *et al.*, 2004).

1. Un grand merci à tous les évaluateurs pour le temps qu'ils ont consacré à l'évaluation et pour leurs commentaires.

2.2. *Scherer*

(Scherer, 1987) considère les émotions comme un processus faisant intervenir plusieurs composantes, dont une composante cognitive. Il introduit un processus d'évaluation consistant en une séquence d'étapes de traitement du stimulus, les *Stimulus Evaluation Checks*. Ce processus évalue successivement : la nouveauté et le caractère inattendu du stimulus, son agréabilité intrinsèque, sa congruence avec les buts de l'individu, les possibilités de *coping* (comment y faire face), et sa compatibilité avec les normes. Contrairement à Lazarus, ces évaluations sont ordonnées. Scherer y associe alors des réponses corporelles, en particulier des expressions faciales en termes d'*Action Units*, éléments définis par (Ekman *et al.*, 2002) pour représenter les mouvements des muscles du visage. Cette théorie est donc bien adaptée pour représenter la dynamique des expressions faciales d'un agent animé (*e.g*. (Grizard *et al.*, 2006)).

2.3. *Frijda*

(Frijda, 1986) insiste plus sur les tendances à l'action induites par les émotions. Un stimulus traverse d'abord plusieurs étapes d'évaluation afin de déterminer ses caractéristiques (causes et conséquences, pertinence et congruence avec les intérêts de l'individu, possibilités de *coping*, urgence). Un signal de contrôle est alors déclenché pour suspendre ou interrompre l'action courante. Une préparation à l'action est générée qui induit ensuite des modifications physiologiques, et finalement une action est sélectionnée et exécutée. (Dastani *et al.*, 2006) s'inspirent de cette notion de tendances à l'action pour définir l'effet des émotions sur les plans d'un agent.

2.4. *Ortony, Clore and Collins*

(Ortony *et al.*, 1988) ont proposé une théorie de l'évaluation cognitive structurée en trois branches correspondant à trois types de stimuli : les conséquences d'événements, les actions d'agents, et les aspects d'objets. Chacun de ces stimuli est évalué suivant un critère central : l'agréabilité d'un événement dépend de la correspondance de ses conséquences avec les buts de l'individu ; l'approbation d'une action dépend de son respect des normes en vigueur ; l'attractivité d'un objet dépend de la correspondance de ses aspects avec les goûts de l'individu. Des critères d'évaluation secondaires influencent l'intensité de l'émotion générée : probabilité d'un événement, responsabilité de l'auteur d'une action, efforts engagés...

Cette typologie regroupe vingt-deux émotions dans huit classes. La première branche regroupe quatre classes d'émotions déclenchées par l'évaluation des conséquences d'un événement. Les émotions de **bien-être** (joie, tristesse) concernent l'évaluation d'un événement en se focalisant uniquement sur sa désirabilité pour l'agent lui-même. Les émotions **par anticipation** (espoir, peur) concernent l'évaluation de la perspective d'un événement en se focalisant sur sa désirabilité pour l'agent lui-même. Les émotions de **confirmation** (déception, soulagement, confirmation de crainte, sa-

tisfaction) concernent l'évaluation de la confirmation ou l'infirmation de la perspective d'un événement en se focalisant sur sa désirabilité pour l'agent lui-même. Les émotions au sujet du **destin d'autrui** (désolé pour, content pour, ressentiment, jubilation) concernent l'évaluation d'un événement en s'intéressant à sa désirabilité pour un autre agent. La deuxième branche regroupe deux classes d'émotions déclenchées par l'évaluation d'une action. Les émotions d'**attribution** (fierté, honte, admiration, reproche) concernent l'évaluation d'une action d'un agent (soi ou autrui) en se focalisant uniquement sur la responsabilité et l'approbation de cette action (et non sur ses conséquences). Les émotions composées **bien-être/attribution** (remords, gratification, gratitude, colère) concernent l'évaluation conjointe de la responsabilité de l'auteur de l'action et de la désirabilité de ses conséquences. La troisième branche contient une classe, les émotions d'**attraction** (haine, amour) déclenchées par l'évaluation de l'attractivité des aspects d'un objet[2].

Il est important de noter que les auteurs de la typologie OCC destinaient celle-ci à une utilisation en intelligence artificielle : « (...) we would like to lay the foundation for a computationally tractable model of emotion. In other words, we would like an account of emotion that could in principle be used in an Artificial Intelligence (AI) system that would, for example, be able to reason about emotion. » (Ortony *et al.*, 1988, p. 2). D'une certaine façon, ce but a été atteint puisque comme nous l'avons dit plus haut, la typologie OCC est actuellement le modèle le plus utilisé pour le développement d'agents émotionnels (de Rosis *et al.*, 2003; Elliott, 1992; Reilly, 1996), même si ce n'est pas le seul (*e.g.* (Gratch *et al.*, 2004) fondés sur (Lazarus, 1991), ou (Staller *et al.*, 2001) fondés sur (Frijda, 1986)).

Les théories de l'évaluation cognitive diffèrent entre elles par les critères d'évaluation retenus, leur ordre d'application, et la définition précise des émotions en fonction de ces critères. Nous avons choisi le modèle OCC pour sa simplicité et sa structuration. Cependant nous avons aussi voulu évaluer la pertinence de ce modèle. Notre agent PLEIAD est donc non seulement un agent émotionnel, mais permet aussi d'évaluer la théorie sous-jacente. Dans la section suivante nous en décrivons les bases théoriques.

3. PLEIAD : modèle théorique

Comme nous l'avons dit en introduction, PLEIAD est un agent émotionnel dont l'architecture a été entièrement formalisée dans une logique modale de type BDI (pour *belief, desire, intention*). Dans cette section, nous présentons ce formalisme.[3]

2. Pour Ortony *et al.*, les objets peuvent englober aussi bien des objets inanimés (par exemple le caviar) que des animaux (par exemple les chiens), des personnes (par exemple sa belle-mère) ou encore des activités (par exemple le basketball).

3. La trame de ce cadre formel a été présentée dans (Adam *et al.*, 2006c). Ce qui suit constitue une version détaillée de la sémantique et de l'axiomatique, avec une notion de désirabilité simplifiée et une notion de temps différente.

3.1. *Architecture BDI*

Notre cadre formel est basé sur la logique modale de (Herzig *et al.*, 2004) qui est un raffinement des travaux de (Cohen *et al.*, 1990). Nous n'utilisons ni le choix ni l'intention. Nous étendons cette logique avec l'opérateur modal de probabilité défini par (Herzig, 2003), ainsi qu'avec des opérateurs de désirabilité et d'obligation.

3.1.1. *Sémantique*

Soit AGT l'ensemble des agents et ACT l'ensemble des actions.

3.1.1.1. Modèle de Kripke

Une sémantique des mondes possibles est utilisée, et un modèle $\mathcal{M}$ est un triplet $\langle W, V, \mathscr{R}\rangle$ où W est un ensemble de mondes possibles, V est une fonction d'interprétation qui associe à chaque monde w l'ensemble V_w des propositions atomiques vraies dans w, et $\mathscr{R}$ est un 6-uplet de structures constitué de :

– $\mathscr{A} : ACT \rightarrow (W \rightarrow 2^W)$ qui, à chaque action $\alpha \in ACT$ et monde possible $w \in W$, associe l'ensemble $\mathscr{A}_\alpha(w)$ des mondes possibles résultant de l'exécution de l'action α dans w. Nous imposons que pour tout $w \in W$, si $w' \in \mathscr{A}_\alpha(w)$ et $w'' \in \mathscr{A}_\beta(w)$ alors $w' = w''$, ce qui impose que les actions sont organisées en histoires, et entraîne qu'elles sont déterministes (il suffit de prendre $\alpha = \beta$) ;

– $\mathscr{B} : AGT \rightarrow (W \rightarrow 2^W)$ qui, à chaque agent $i \in AGT$ et monde possible $w \in W$, associe l'ensemble $\mathscr{B}_i(w)$ des mondes possibles compatibles avec les croyances de i dans w. Ces relations d'accessibilité sont sérielles, transitives et euclidiennes ;

– $\mathscr{P} : AGT \rightarrow (W \rightarrow 2^{2^W})$ qui, à chaque agent $i \in AGT$ et monde possible $w \in W$, associe un ensemble d'ensembles de mondes possibles $\mathscr{P}_i(w)$. Suivant (Chellas, 1980, chap. 8), ces ensembles de mondes possibles sont appelés des *voisinages*. Nous formalisons les voisinages comme des sous-ensembles de $\mathscr{B}_i(w)$: dans ces conditions, tout $U \in \mathscr{P}_i(w)$ contient intuitivement plus d'éléments que son complémentaire $\mathscr{B}_i(w) \setminus U$. Bien que la contrainte suivante aille dans ce sens, elle est plus faible [4] et ne capture pas complètement cette intuition : pour tout $w \in W$, si $U_1, U_2 \in \mathscr{P}_i(w)$ alors $U_1 \cap U_2 \neq \emptyset$. Autrement dit, si φ est probable (*i.e.* φ est vrai dans tous les mondes d'un voisinage), alors $\neg\varphi$ ne l'est pas (chacun des autres voisinages contiendra au moins un monde où φ est vrai). Enfin, afin de garantir qu'au moins les tautologies soient probables, nous imposons que $\mathscr{P}_i(w) \neq \emptyset$ pour tout $w \in W$;

– $\mathscr{D} : AGT \rightarrow (W \rightarrow 2^W)$ qui, à chaque agent $i \in AGT$ et monde possible $w \in W$, associe l'ensemble $\mathscr{D}_i(w)$ des mondes idéaux compatibles avec ce qui est désirable pour l'agent i dans le monde w. Ces relations d'accessibilité $\mathscr{D}_i$ sont sérielles ;

– $\mathscr{G} : W \rightarrow 2^W$ qui, à chaque monde possible $w \in W$, associe l'ensemble $\mathscr{G}(w)$ des mondes possibles du futur de w. Cette relation est un ordre linéaire (réflexive, transitive and antisymétrique).

4. On peut exhiber sur un exemple une distribution de voisinages satisfaisant les contraintes énoncées mais « collectant » moins de 50 % des mondes, *cf.* (Walley *et al.*, 1979).

– $\mathscr{I} : AGT \rightarrow (W \rightarrow 2^W)$ qui, à chaque agent $i \in AGT$ et monde possible $w \in W$, associe l'ensemble $\mathscr{I}_i(w)$ des mondes idéaux pour l'agent i. Dans ces mondes idéaux toutes les obligations, normes, standards... (sociaux, légaux, moraux...) de l'agent i sont respectés. Toutes ces relations sont sérielles.

De plus, nous imposons des contraintes entre certaines relations :

– si $w \in \mathscr{B}_i(w')$ alors $\mathscr{P}_i(w) = \mathscr{P}_i(w')$ et $\mathscr{D}_i(w) = \mathscr{D}_i(w')$, ce qui garantit que les agents sont conscients de ce qu'ils croient probable ainsi que de leurs désirs[5] ;

– $U \subseteq \mathscr{B}_i(w)$ pour tout $U \in \mathscr{P}_i(w)$, ainsi la croyance implique la probabilité ;

– $\mathscr{G} \supseteq \mathscr{A}_\alpha$ pour tout α, ce qui garantit que le futur de w contient les mondes résultant de l'exécution d'actions dans w. Par ailleurs, si $\mathscr{G} \supseteq \mathscr{A}_\alpha$ et $w' \in \mathscr{A}_\alpha^{-1}(w)$, *i.e.* $w \in \mathscr{A}_\alpha(w')$, alors $w \in \mathscr{G}(w')$, d'où $w' \in \mathscr{G}^{-1}(w)$. En d'autres termes, $\mathscr{G}^{-1} \supseteq \mathscr{A}_\alpha^{-1}$, ce qui garantit une propriété similaire dans le passé ;

– enfin, par souci de simplicité, nous faisons l'hypothèse que ce que l'agent aime persiste : si $w\mathscr{G}w'$ alors $\mathscr{D}_i(w) = \mathscr{D}_i(w')$;

– de même pour les obligations, normes, standards... (sociaux, légaux, moraux...) auxquels les agents sont soumis : si $w\mathscr{G}w'$ alors $\mathscr{I}_i(w) = \mathscr{I}_i(w')$.

Nous sommes conscients qu'en toute généralité cette dernière contrainte est trop forte. Néanmoins, elle apparaît assez réaliste dans le cas où les intervalles temporels considérés sont relativement courts (par exemple dans le cas d'un petit dialogue).

3.1.1.2. Opérateurs modaux et langage

Nous associons des opérateurs modaux à ces relations d'accessibilité :

– $After_\alpha\, \varphi$ signifie que « φ est vraie après toute exécution de l'action α » ;

– $Before_\alpha\, \varphi$ signifie que « φ est vraie avant toute exécution de l'action α » ;

– $Bel_i\, \varphi$ signifie que « l'agent i croit que φ » ;

– $Prob_i\, \varphi$ signifie que « pour i φ est plus probable que $\neg\varphi$ » ;

– $Des_i\, \varphi$ signifie que « φ est désirable pour i » ;

– $G\varphi$ signifie que « désormais φ est vraie » ;

– $H\varphi$ signifie que « φ a toujours été vraie dans le passé » ;

– $Idl_i\, \varphi$ signifie que « idéalement pour i il est le cas que φ ».

L'opérateur H est le réciproque de G, et $Before_\alpha$ celui de $After_\alpha$.

$ATM = \{p, q...\}$ est l'ensemble des formules atomiques de la logique classique. Toute formule atomique est une formule complexe. Si φ et ψ sont deux formules complexes et $\Box$ un des opérateurs modaux ci-dessus, alors $\neg\varphi$, $\varphi \vee \psi$ et $\Box\varphi$ sont des formules complexes. L'ensemble des formules complexes est $FORM = \{\varphi, \psi...\}$.

5. En raison de la transitivité et de l'euclidianité des relations $\mathscr{B}_i$, ils sont également conscients de leurs croyances, *i.e.* on peut dériver la propriété suivante : si $w \in \mathscr{B}_i(w')$ alors $\mathscr{B}_i(w) = \mathscr{B}_i(w')$.

Si φ et ψ sont des formules complexes, on définit de la manière usuelle $\varphi \wedge \psi$ et $\varphi \rightarrow \psi$ comme étant des abréviations de formules complexes. Nous définissons également les abréviations suivantes :

– $Happens_\alpha\,\varphi \stackrel{déf}{=} \neg After_\alpha\,\neg\varphi$: « α va être exécutée, après quoi φ » ;

– $Done_\alpha\,\varphi \stackrel{déf}{=} \neg Before_\alpha\,\neg\varphi$: « α vient d'être exécutée, et φ était vraie avant » ;

– $Undes_i\,\varphi \stackrel{déf}{=} Des_i\,\neg\varphi$ signifie que φ est indésirable pour l'agent i[6] ;

– $F\varphi \stackrel{déf}{=} \neg G\neg\varphi$: « φ est vraie ou sera vraie à un instant dans le futur » ;

– $P\varphi \stackrel{déf}{=} \neg H\neg\varphi$ signifie que « φ est ou a été vraie ».

3.1.1.3. Conditions de vérité

Les conditions de vérité sont standards pour presque tous nos opérateurs :

$$w \Vdash \Box\varphi \quad \text{ssi} \quad w' \Vdash \varphi \text{ pour tout } w' \in \mathscr{R}_\Box(w)$$

où $(\Box, \mathscr{R}_\Box)$ est soit $(After_\alpha, \mathscr{A}_\alpha)$ avec $\alpha \in ACT$, soit $(Bel_i, \mathscr{B}_i)$ avec $i \in AGT$, soit $(Des_i, \mathscr{D}_i)$ avec $i \in AGT$, soit $(G, \mathscr{G})$, soit $(Idl_i, \mathscr{I}_i)$ avec $i \in AGT$.

Pour les opérateurs réciproques nous avons :

$$w \Vdash \boxminus\varphi \quad \text{ssi} \quad w' \Vdash \varphi \text{ pour tout } w' \text{ tel que } w \in \mathscr{R}_\boxminus(w')$$

où $(\boxminus, \mathscr{R}_\boxminus)$ est soit $(Before_\alpha, \mathscr{A}_\alpha)$ avec $\alpha \in ACT$, soit $(H, \mathscr{G})$.

De plus :

$$w \Vdash Prob_i\,\varphi \quad \text{ssi} \quad \text{il existe } U \in \mathscr{P}_i(w) \text{ tel que pour tout } w' \in U,\ w' \Vdash \varphi.$$

3.1.2. *Axiomatique*

3.1.2.1. L'action

$After_\alpha$ et $Before_\alpha$ sont définis dans la logique standard du temps (*tense logic*) $\mathbf{K}_t$, *i.e.* une logique normale[7] **K** étendue avec les axiomes de conversion suivants (*cf.* (Burgess, 2002) pour plus de détails) :

$$Happens_\alpha\,\varphi \rightarrow After_\beta\,\varphi \qquad \text{[CD-ACT]}$$

$$\varphi \rightarrow After_\alpha\,Done_\alpha\,\varphi \qquad \text{[CONV-AD]}$$

$$\varphi \rightarrow Before_\alpha\,Happens_\alpha\,\varphi \qquad \text{[CONV-BH]}$$

6. Dans (Adam *et al.*, 2006c) les opérateurs d'indésirabilité étaient définis à partir de deux opérateurs normaux et entretenaient avec les opérateurs de désirabilité une relation bipolaire : quelque chose de désirable n'induisait pas nécessairement que son contraire était indésirable.

7. $\Box$ est un opérateur normal si et seulement si l'axiome [K-$\Box$] : $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$ et la règle de nécessitation [RN-$\Box$] : $\frac{\varphi}{\Box\varphi}$ sont valides. *Cf.* (Chellas, 1980, chap. 4) pour le détail des propriétés formelles des logiques normales.

[CD-ACT] signifie que les actions sont organisées en histoires[8], et implique que les actions sont déterministes (on le voit en posant $\alpha = \beta$). [CONV-AD] et [CONV-BH] rendent compte du lien entre passé et futur.

Dans la suite, la notation $i{:}\alpha$ se lit « l'agent i est l'auteur de l'action α ».

3.1.2.2. La croyance

Les opérateurs Bel_i sont définis dans la logique standard **KD45** (Hintikka, 1962). Les axiomes correspondants sont ceux des logiques normales plus les suivants :

$$Bel_i\,\varphi \rightarrow \neg Bel_i\,\neg\varphi \qquad [\text{D-}Bel_i]$$

$$Bel_i\,\varphi \rightarrow Bel_i\,Bel_i\,\varphi \qquad [\text{4-}Bel_i]$$

$$\neg Bel_i\,\varphi \rightarrow Bel_i\,\neg Bel_i\,\varphi \qquad [\text{5-}Bel_i]$$

Ainsi les croyances d'un agent sont consistantes [D-Bel_i], et un agent est conscient de ce qu'il croit [4-Bel_i] et de ce qu'il ne croit pas [5-Bel_i].

3.1.2.3. Le temps

Les opérateurs G et H sont définis dans la logique du temps linéaire $\mathbf{S4.3}_t$ (Burgess, 2002) qui correspond à une logique normale **K** pour chacun des opérateurs plus les axiomes suivants :

$$G\varphi \rightarrow \varphi \qquad [\text{T-}G]$$

$$(F\varphi \wedge F\psi) \rightarrow F(\varphi \wedge F\psi) \vee F(\psi \wedge F\varphi) \qquad [\text{3-}F]$$

$$G\varphi \rightarrow GG\varphi \qquad [\text{4-}G]$$

$$H\varphi \rightarrow \varphi \qquad [\text{T-}H]$$

$$(P\varphi \wedge P\psi) \rightarrow P(\varphi \wedge P\psi) \vee P(\psi \wedge P\varphi) \qquad [\text{3-}P]$$

$$H\varphi \rightarrow HH\varphi \qquad [\text{4-}H]$$

$$\varphi \rightarrow GP\varphi \qquad [\text{CONV-GP}]$$

$$\varphi \rightarrow HF\varphi \qquad [\text{CONV-HF}]$$

[T-G] et [T-H] signifient que futur et passé sont des notions prises au sens large : si une proposition est toujours vraie dans le futur ou le passé, elle l'est en particulier dans l'instant présent. [CONV-GP] et [CONV-HF] rendent compte du lien entre passé et futur. [4-G] et [4-H] expriment la transitivité du temps dans le futur et le passé: si φ est vraie dans tous les futurs (resp. dans tous les passés), alors φ est aussi vraie dans tous les futurs de tous ces futurs (resp. dans tous les passés de tous ces passés). [3-F] et [3-P] indiquent que si deux formules sont vraies à deux instants (pas forcément

8. Cela n'empêche pas l'exécution parallèle de plusieurs actions, mais garantit qu'elles mènent alors toutes vers le même monde.

identiques) dans le futur (respec. dans le passé) alors l'une est forcément vraie avant l'autre. Cela induit que le temps est linéaire dans le futur et le passé. Cela peut sembler critiquable dans la mesure où intuitivement le futur est arborescent. En fait, comme c'est la perception que des agents ont du temps plutôt que la nature du temps qui nous importe, la diversité des futurs peut être représentée par des histoires différentes dont les présents sont des mondes épistémiques différents (*cf.* figure 1).

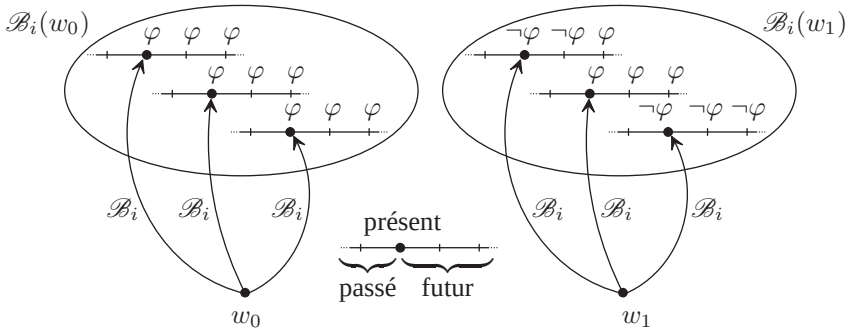

Figure 1. *Représentation du temps linéaire où* $w_0 \Vdash Bel_i\, G\varphi$, $w_1 \Vdash \neg Bel_i\, \neg(\neg\varphi \wedge F\varphi)$, $w_1 \Vdash \neg Bel_i\, \neg G\varphi$ *et* $w_1 \Vdash \neg Bel_i\, \neg G\neg\varphi$

3.1.2.4. La probabilité

Les opérateurs de probabilité correspondent à la notion de croyance faible, basée sur la notion de mesure de probabilité subjective (ce qui est capturé sémantiquement par le fait que les mondes probables appartiennent à l'ensemble des mondes crus).

La logique de $Prob$ est plus faible que celle de la croyance. En particulier, l'ensemble des probabilités n'est pas fermé sous la conjonction, *i.e.* l'axiome [C-$Prob_i$] : $(Prob_i\, \varphi \wedge Prob_i\, \psi) \rightarrow Prob_i\, (\varphi \wedge \psi)$ n'est pas valide, ce qui suffit à en faire une logique non-normale (Chellas, 1980, Théorème 4.3).

La condition de vérité et les conditions sémantiques valident ces principes :

$$\frac{\varphi \rightarrow \psi}{Prob_i\, \varphi \rightarrow Prob_i\, \psi} \qquad [\text{RM-}Prob_i]$$

$$Prob_i\, \top \qquad [\text{N-}Prob_i]$$

$$Prob_i\, \varphi \rightarrow \neg Prob_i\, \neg\varphi \qquad [\text{D-}Prob_i]$$

Ces trois principes caractérisent l'ensemble des probabilités : [RM-$Prob_i$] signifie qu'il est fermé sous la conséquence logique ; [N-$Prob_i$] signifie qu'il contient les tautologies ; [D-$Prob_i$] signifie qu'il est consistant.

À partir de [RM-$Prob_i$] et [N-$Prob_i$] on peut prouver la règle d'inférence suivante (il suffit de prendre $\varphi = \top$ dans [RM-$Prob_i$]) :

$$\frac{\varphi}{Prob_i\,\varphi} \qquad \text{[RN-}Prob_i\text{]}$$

Enfin, nous avons des principes d'introspection pour $Prob_i$ (*cf.* parag. 3.1.2.7).

3.1.2.5. La désirabilité

Sa sémantique est identique à celle de la logique déontique standard (SDL) et s'exprime aussi en termes de mondes idéaux : Des_i et $Undes_i$ entretiennent la même relation que l'obligation et l'interdiction. Ainsi, la logique associée à l'opérateur Des_i est **KD**, *i.e.* la logique normale **K** plus l'axiome suivant qui rend les désirs consistants :

$$Des_i\,\varphi \rightarrow \neg Des_i\,\neg\varphi \qquad \text{[D-}Des_i\text{]}$$

Ce qui est désirable pour un agent représente des préférences individuelles. Celles-ci peuvent être irréalistes car nous n'imposons pas que $\mathscr{B}_i(w) \cap \mathscr{D}_i(w) \neq \emptyset$: un agent peut désirer être dans des états qu'il croit actuellement impossibles.

Nous soulignons que les désirs ne sont fermés ni sous l'implication ni sous la conjonction : je peux désirer épouser Anne et épouser Beth sans désirer être bigame.

3.1.2.6. L'idéalité

L'idéalité est vue ici comme une notion d'obligation prise dans un sens très large : elle inclut toutes les règles provenant de l'extérieur de l'agent que celui-ci doit idéalement respecter. Ces règles peuvent être explicites (comme des lois) ou plus ou moins implicites (comme les obligations morales ou sociales).

La logique de l'idéalité est donc tout naturellement identique à SDL, *i.e.* la logique normale **K** plus l'axiome suivant qui rend les idéaux consistants :

$$Idl_i\,\varphi \rightarrow \neg Idl_i\,\neg\varphi \qquad \text{[D-}Idl_i\text{]}$$

3.1.2.7. Axiomes mixtes

Des contraintes sémantiques stipulent que certains opérateurs modaux sont interdépendants. Les agents sont conscients de ce qu'ils croient probable et de leurs désirs :

$$Prob_i\,\varphi \rightarrow Bel_i\,Prob_i\,\varphi \qquad \text{[4-MIX1]}$$

$$\neg Prob_i\,\varphi \rightarrow Bel_i\,\neg Prob_i\,\varphi \qquad \text{[5-MIX1]}$$

$$Des_i\,\varphi \rightarrow Bel_i\,Des_i\,\varphi \qquad \text{[4-MIX2]}$$

$$\neg Des_i\,\varphi \rightarrow Bel_i\,\neg Des_i\,\varphi \qquad \text{[5-MIX2]}$$

À partir de ces axiomes et de [D-$Prob_i$] et [D-Des_i], on peut facilement prouver leur réciproque et obtenir ainsi des équivalences.

Les mondes probables sont inclus dans l'ensemble des mondes épistémiques :

$$(Bel_i\,\varphi \wedge Prob_i\,\psi) \rightarrow Prob_i\,(\varphi \wedge \psi) \qquad \text{[C-MIX]}$$

Cet axiome permet de dériver les théorèmes suivants :

$$Bel_i\,\varphi \rightarrow Prob_i\,\varphi \qquad [1]$$

$$Prob_i\,\varphi \rightarrow \neg Bel_i\,\neg\varphi \qquad [2]$$

Par ailleurs, le temps et l'action sont liés : si φ est tout le temps vrai dans le futur alors après toute exécution d'une action φ sera vrai. De même, si φ est tout le temps vrai dans le passé alors avant toute exécution d'une action φ était vrai. Soit :

$$G\varphi \rightarrow After_{\alpha}\,\varphi \qquad \text{[GA-MIX]}$$

$$H\varphi \rightarrow Before_{\alpha}\,\varphi \qquad \text{[HB-MIX]}$$

Enfin, comme nous l'avons dit et justifié lors de la présentation de notre sémantique, nous considérons les désirs, les non-désirs et les idéaux imposés à l'agent comme persistants (*i.e.* ils sont préservés à travers le temps et l'action). Soit :

$$Des_i\,\varphi \rightarrow G Des_i\,\varphi \qquad \text{[Pers-}Des_i\text{]}$$

$$\neg Des_i\,\varphi \rightarrow G\neg Des_i\,\varphi \qquad \text{[Pers-}\neg Des_i\text{]}$$

$$Idl_i\,\varphi \rightarrow G Idl_i\,\varphi \qquad \text{[Pers-}Idl_i\text{]}$$

$$\neg Idl_i\,\varphi \rightarrow G\neg Idl_i\,\varphi \qquad \text{[Pers-}\neg Idl_i\text{]}$$

Notons que nous n'avons pas postulé que les agents sont conscients des standards. En effet, intuitivement, il serait trop fort d'avoir les principes $Idl_i\,\varphi \rightarrow Bel_i\,Idl_i\,\varphi$ et $\neg Idl_i\,\varphi \rightarrow Bel_i\,\neg Idl_i\,\varphi$, et ceux-ci ne sont donc pas valides dans notre sémantique.

3.2. *Formalisation des émotions*

Nous utilisons maintenant ce formalisme logique pour exprimer les émotions du modèle OCC. Nous donnons pour chaque émotion sa définition par Ortony *et al.* suivie de notre définition logique (*cf.* (Adam *et al.*, 2006c) pour plus de détails).

Afin de clarifier les concepts utilisés, nous utiliserons l'abréviation suivante :

$$Expect_i\,\varphi \stackrel{déf}{=} Prob_i\,\varphi \wedge \neg Bel_i\,\varphi \qquad \text{[Déf-}Expect_i\text{]}$$

qui représente le fait que l'agent i s'attend plutôt à ce que φ soit vrai, bien qu'il envisage la possibilité qu'il soit faux. (Notons que $Expect_i\,\varphi \rightarrow \neg Bel_i\,\neg\varphi$, *i.e.* l'attente implique nécessairement le fait que l'agent envisage la possibilité que φ soit vrai.)

3.2.1. *Émotions concernant des conséquences d'événements*

Les conditions de déclenchement de ces émotions dépendent de la désirabilité de l'événement évalué (qui peut être immédiat ou anticipé), c'est-à-dire de son impact sur les buts de l'agent considéré (soi-même ou autrui). Nous pouvons définir :

– deux émotions concernant des événements qui se sont produits : un agent ressent de la joie (resp. de la tristesse) s'il est content (resp. mécontent) à propos d'un événement désirable (resp. indésirable) pour lui-même :

$$Joy_i\,\varphi \stackrel{déf}{=} Bel_i\,\varphi \wedge Des_i\,\varphi$$

$$Distress_i\,\varphi \stackrel{déf}{=} Bel_i\,\varphi \wedge Undes_i\,\varphi$$

– deux émotions concernant des perspectives d'événements : un agent ressent de l'espoir (resp. de la crainte) s'il est content (resp. mécontent) à propos de la perspective d'un événement désirable (resp. indésirable) pour lui :

$$Hope_i\,\varphi \stackrel{déf}{=} Expect_i\,\varphi \wedge Des_i\,\varphi$$

$$Fear_i\,\varphi \stackrel{déf}{=} Expect_i\,\varphi \wedge Undes_i\,\varphi$$

– deux émotions concernant la confirmation d'une perspective d'événement antérieur : un agent ressent de la confirmation de crainte (resp. de la satisfaction) s'il est mécontent (resp. content) à propos de la confirmation d'un événement indésirable (resp. désirable) pour lui :

$$FearConfirmed_i\,\varphi \stackrel{déf}{=} Bel_i\,PExpect_i\,\varphi \wedge Undes_i\,\varphi \wedge Bel_i\,\varphi$$

$$Satisfaction_i\,\varphi \stackrel{déf}{=} Bel_i\,PExpect_i\,\varphi \wedge Des_i\,\varphi \wedge Bel_i\,\varphi$$

– deux émotions concernant la non-confirmation d'une perspective d'événement : un agent ressent du soulagement (resp. de la déception) s'il est content (resp. mécontent) à propos de la non confirmation de la perspective d'un événement indésirable (resp. désirable) pour lui :

$$Relief_i\,\varphi \stackrel{déf}{=} Bel_i\,PExpect_i\,\neg\varphi \wedge Undes_i\,\neg\varphi \wedge Bel_i\,\varphi$$

$$Disappointment_i\,\varphi \stackrel{déf}{=} Bel_i\,PExpect_i\,\neg\varphi \wedge Des_i\,\neg\varphi \wedge Bel_i\,\varphi$$

– deux émotions concernant un événement touchant un ami de l'agent : un agent est heureux pour (resp. désolé pour) un autre agent s'il est content (resp. mécontent) à propos d'un événement qu'il présume désirable (resp. indésirable) pour cet agent :

$$HappyFor_{i,j}\varphi \stackrel{déf}{=} Bel_i\,\varphi \wedge Bel_i\,F\,Bel_j\,\varphi \wedge Bel_i\,Des_j\,\varphi \wedge Des_i\,Bel_j\,\varphi$$

$$SorryFor_{i,j}\varphi \stackrel{déf}{=} Bel_i\,\varphi \wedge Bel_i\,F\,Bel_j\,\varphi \wedge Bel_i\,Undes_j\,\varphi \wedge Undes_i\,Bel_j\,\varphi$$

– deux émotions concernant un événement touchant un ennemi de l'agent : un agent ressent du ressentiment (resp. de la jubilation) envers un autre agent s'il est mécontent (resp. content) à propos de la survenue d'un événement qu'il présume désirable (resp. indésirable) pour cet agent :

$$Resentment_{i,j}\varphi \stackrel{déf}{=} Bel_i\,\varphi \wedge Bel_i\,F Bel_j\,\varphi \wedge Bel_i\,Des_j\,\varphi \wedge Undes_i\,Bel_j\,\varphi$$

$$Gloating_{i,j}\varphi \stackrel{déf}{=} Bel_i\,\varphi \wedge Bel_i\,F Bel_j\,\varphi \wedge Bel_i\,Undes_j\,\varphi \wedge Des_i\,Bel_j\,\varphi$$

3.2.2. *Émotions concernant des actions*

Les conditions de déclenchement des émotions concernant les actions dépendent du caractère méritoire ou répréhensible de l'action évaluée, c'est-à-dire du respect ou non des standards par son auteur (qui est éventuellement différent de l'agent qui ressent l'émotion). Nous utilisons donc notre opérateur déontique Obl pour définir :

– deux émotions concernant une action de l'agent lui-même : un agent ressent de la fierté (resp. de la honte) s'il approuve (resp. désapprouve) sa propre action méritoire (resp. répréhensible)

$$\begin{aligned} Pride_i\,i{:}\alpha \stackrel{déf}{=} & Bel_i\,Done_{i:\alpha}\,(\neg Prob_i\,Happens_{i:\alpha}\,\top \wedge \\ & Bel_i\,Obl_i\,Happens_{i:\alpha}\,\top) \end{aligned}$$

$$\begin{aligned} Shame_i\,i{:}\alpha \stackrel{déf}{=} & Bel_i\,Done_{i:\alpha}\,(\neg Prob_i\,Happens_{i:\alpha}\,\top \wedge \\ & Bel_i\,Obl_i\,\neg Happens_{i:\alpha}\,\top) \end{aligned}$$

– deux émotions concernant une action d'un autre agent : un agent ressent de l'admiration (resp. du reproche) envers un autre agent s'il approuve (resp. désapprouve) une action méritoire (resp. répréhensible) de cet agent

$$\begin{aligned} Admiration_{i,j}\,j{:}\alpha \stackrel{déf}{=} & Bel_i\,Done_{j:\alpha}\,(\neg Prob_i\,Happens_{j:\alpha}\,\top \wedge \\ & Bel_i\,Obl_j\,Happens_{j:\alpha}\,\top) \end{aligned}$$

$$\begin{aligned} Reproach_{i,j}\,j{:}\alpha \stackrel{déf}{=} & Bel_i\,Done_{j:\alpha}\,(\neg Prob_i\,Happens_{j:\alpha}\,\top \wedge \\ & Bel_i\,Obl_j\,\neg Happens_{j:\alpha}\,\top) \end{aligned}$$

– deux émotions concernant une action de l'agent lui-même *et* ses conséquences pour lui-même : selon Ortony et col., la gratification est un mélange de fierté et de joie, et le remords est un mélange de honte et de tristesse

$$Gratification_i\,(i{:}\alpha, \varphi) \stackrel{déf}{=} Pride_i\,i{:}\alpha \wedge Bel_i\,Before_{i:\alpha}\,\neg Bel_i\,F\varphi \wedge Joy_i\,\varphi$$

$$Remorse_i\,(i{:}\alpha, \varphi) \stackrel{déf}{=} Shame_i\,i{:}\alpha \wedge Bel_i\,Before_{i:\alpha}\,\neg Bel_i\,F\varphi \wedge Distress_i\,\varphi$$

– deux émotions concernant une action d'un autre agent *et* ses conséquences pour l'agent qui ressent l'émotion : selon Ortony et col., la gratitude est un mélange d'admiration et de joie, et la colère est un mélange de reproche et de tristesse

$$\begin{aligned} Gratitude_{i,j}(j{:}\alpha, \varphi) &\stackrel{déf}{=} Admiration_{i,j}\, j{:}\alpha \wedge \\ &\quad Bel_i\, Before_{j:\alpha}\, \neg Bel_i\, F\varphi \wedge Joy_i\, \varphi \\ Anger_{i,j}(j{:}\alpha, \varphi) &\stackrel{déf}{=} Reproach_{i,j}\, j{:}\alpha \wedge \\ &\quad Bel_i\, Before_{j:\alpha}\, \neg Bel_i\, F\varphi \wedge Distress_i\, \varphi \end{aligned}$$

3.2.3. *Conscience des émotions*

Notre logique, grâce aux axiomes d'introspection pour tous les opérateurs, nous permet de démontrer que l'agent est conscient de ses émotions, c'est-à-dire que $Emotion_i\varphi \leftrightarrow Bel_i\, Emotion_i\varphi$ et $\neg Emotion_i\varphi \leftrightarrow Bel_i\, \neg Emotion_i\varphi$ sont valides pour toute $Emotion_i$ parmi les vingt émotions définies ci-dessus.

4. PLEIAD : implémentation et évaluation

4.1. *Implémentation*

PLEIAD est un agent émotionnel fondé sur une architecture BDI des émotions entièrement formalisée[9]. Par rapport à la théorie d'origine nous avons apporté quelques modifications. La première a consisté à associer des degrés aux attitudes mentales de l'agent pour en déduire une intensité associée à chaque émotion. Ces valeurs numériques n'étaient pas présentes dans notre cadre logique car il est très difficile d'y associer une sémantique (*e.g.* (Laverny *et al.*, 2005)). Cependant elles permettent indéniablement d'augmenter le réalisme expressif de l'agent (*cf.* partie 4.1.4).

La deuxième modification est une concession en vue de simplifier l'implémentation, et concerne la non-complétude du démonstrateur : celui-ci est implémenté en Prolog, qui n'est pas complet ; de plus, nous ne gérons pas tous les connecteurs logiques (les attitudes mentales portent principalement sur des formules atomiques) et la sémantique de la logique n'est pas totalement implémentée (nous nous sommes limités aux axiomes nous paraissant les plus utiles, et n'avons pas intégré les axiomes d'introspection pour éviter les problèmes de bouclage). Cependant, dans le cadre du fonctionnement normal de notre agent, ce démonstrateur simple s'est avéré suffisant.

4.1.1. *Interface de PLEIAD*

L'interface de PLEIAD permet de créer un agent en remplissant sa base de connaissances de croyances, désirs, normes, attentes... L'utilisateur peut ensuite envoyer un

9. Il a été présenté pour la première fois lors du workshop WACA'2006, *cf.* (Adam, 2006).

stimulus à l'agent (action ou événement) ou modifier ses centres d'attention pour observer sa réaction. Enfin il peut obtenir grâce aux différentes options du menu toutes les informations utiles sur l'agent : attitudes mentales, contenu du focus, émotion exprimée et émotions non exprimées, mémoire, relations avec les autres agents...

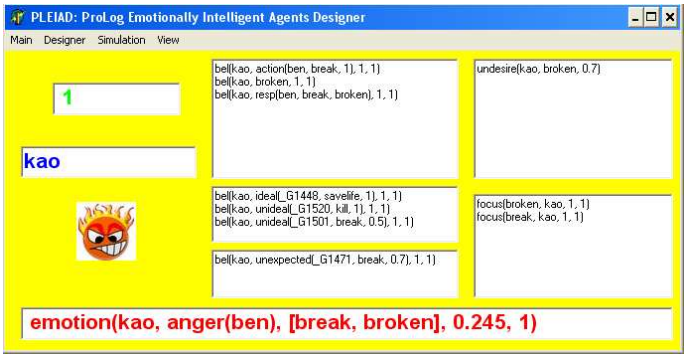

Figure 2. *PLEIAD : interface*

4.1.2. *Architecture de l'agent*

Notre agent dispose d'une base de connaissances (BC) contenant des attitudes mentales (AM) graduées (cf. partie 4.1.4) et associées à un degré d'activation (ou *focus*, cf. partie 4.1.3). L'utilisateur peut envoyer des stimuli à l'agent (actions ou événements) en spécifiant leurs effets, qui sont alors ajoutés directement aux croyances de l'agent. Le module de perception est donc transparent pour l'instant, dans le sens où chaque événement est perçu correctement et entièrement[10]. Un démonstrateur logique est chargé de saturer la BC à tout moment avec toutes les AM déductibles de l'application d'un ensemble d'axiomes sur son contenu. Le module de gestion de l'activation génère les modifications automatiques du focus comme sa décroissance temporelle. Le module émotionnel utilise les définitions formelles des émotions présentées dans la partie précédente pour calculer toutes les émotions déductibles de la BC (*i.e.* les émotions « ressenties » par l'agent selon la théorie OCC), chacune associée à une intensité. Finalement un module d'expression détermine laquelle de ces émotions sera exprimée préférentiellement[11], cette émotion étant destinée à être envoyée à un module d'animation faciale et/ou corporelle que nous ne gérons pas ici. Nous exprimons l'émotion par une ligne de texte associée à un émoticône.

10. Plus tard, nous pourrons envisager une influence des émotions de l'agent sur ses capacités de perception, et donc une perception incomplète ou déformée des stimuli reçus.
11. Le module d'expression pourrait aussi recevoir un vecteur émotionnel afin d'afficher une expression faciale mixte (Ochs *et al.*, 2005).

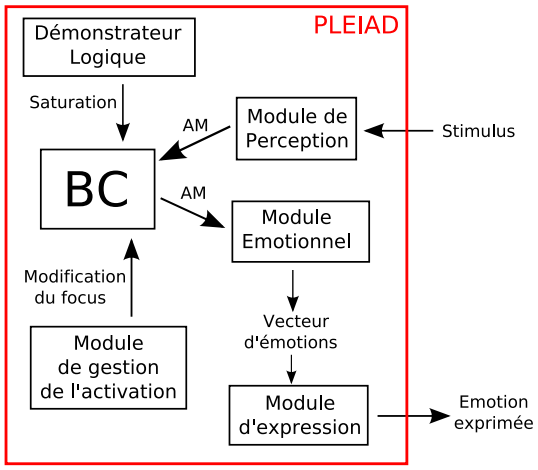

Figure 3. *Architecture de PLEIAD*

4.1.3. *Le module de gestion de l'activation*

(Anderson, 1990) propose une théorie de l'activation, notion qui détermine selon lui l'accessibilité d'une unité de connaissance, ou *chunk*. L'activation d'un *chunk* à un instant donné est déterminé en fonction des expériences passées, proportionnellement à son utilité supposée à cet instant. Il est calculé comme la somme d'une activation de base et d'une activation associative. L'activation de base représente l'utilité du *chunk* dans le passé, *i.e*. sa récence et sa fréquence d'utilisation ; elle décroît logarithmiquement au cours du temps. L'activation associative représente la pertinence du *chunk* dans le contexte actuel et dépend des activations des *chunks* associés[12]. Cette théorie est implémentée dans l'architecture cognitive ACT-R (Lebiere *et al.*, 1993).

Dans PLEIAD, nous associons aux AM contenues dans la BC d'un agent (croyances, désirs...) un degré de focus, qui est une notion simplifiée de l'activation au sens d'Anderson. En effet le focus ne fait pas intervenir de réseau d'association entre propositions, sa valeur est donc constituée uniquement de ce qu'Anderson appelle l'activation de base. Le focus représente l'accessibilité d'une AM pour les processus cognitifs. Dans notre implémentation, le focus concerne les propositions (p), et toutes les AM concernant la même proposition (par exemple $Bel_i\, p$ et $Des_i\, p$) sont également accessibles. Un degré de focus empirique est attribué au départ aux propositions de la BC. De plus, les stimuli reçus par l'agent reçoivent immédiatement un degré de focus maximal (égal à 1). Comme dans ACT-R, ce degré de focus décroît ensuite au cours du temps selon un facteur fixé empiriquement.

12. Un *chunk* est relié à d'autres *chunks* s'il a été nécessaire quand ces autres *chunks* faisaient partie d'un certain but à atteindre.

Nous pouvons ainsi formaliser deux stratégies de raisonnement simples, au sens de Forgas dans ses travaux sur l'*Affect Infusion Model* (Forgas, 1995). Selon lui l'individu choisit une stratégie de raisonnement en fonction du contexte et de manière à minimiser ses efforts. Nous proposons donc deux stratégies pour notre agent : soit il utilise la totalité de sa BC pour raisonner, ce qui est plus efficace mais plus coûteux ; soit il n'utilise que les cognitions les plus activées (celles dont le degré de focus dépasse un certain seuil) donc les plus accessibles, afin d'obtenir une solution plus rapidement. Cependant notre modèle n'explique pas comment l'agent choisit l'une ou l'autre de ces stratégies, et notre agent utilise toujours la totalité de sa BC pour raisonner.

4.1.4. *Calcul de l'intensité des émotions*

Pour augmenter le réalisme des émotions exprimées par l'agent, nous avons tenu à leur faire correspondre une intensité graduelle. En effet pour les humains, être un peu agacé n'est pas du tout la même chose que d'être vraiment en colère. De plus l'importance de l'intensité des émotions, que ce soit pour leur expression ou leur influence sur le comportement, a déjà été soulignée auparavant en même temps que le manque de travaux à leur sujet : « One of the more curious aspects of emotion research indeed is its lack of attention to the fact that emotions vary in intensity. The failure of emotion theorists to address questions concerning emotion intensity is all the more puzzling because intensity is such a salient feature of emotions. Our phenomenal experience acknowledges this fact, as does our behavior and our language; so how is it that our science essentially ignores it? And ignore it, it does. » (Frijda *et al.*, 1992)

Pour (Castelfranchi *et al.*, 2003), le degré et la dynamique des émotions qui émergent de la composition d'attitudes mentales dépend strictement du degré et de la dynamique de ces attitudes mentales. Ils associent donc un degré de certitude subjective aux croyances, et un degré d'importance aux buts de l'agent, pour en déduire le degré de différentes émotions relatives aux attentes. Nous associons donc nous aussi des degrés aux attitudes mentales de l'agent et calculons alors le degré d'intensité d'une émotion comme un produit des degrés des AM intervenant dans sa définition, ainsi que du degré d'activation du concept sur lequel elle porte. Ainsi l'intensité de l'émotion décroît naturellement à mesure que l'activation de son objet diminue.

Les formules de logique BDI décrivant la définition des émotions ont été traduites en Prolog en ajoutant le calcul de l'intensité. Par exemple pour l'émotion de joie :

– Formule BDI : $Joy_a\,\varphi \stackrel{déf}{=} Bel_a\,\varphi \wedge Des_a\,\varphi$

– Prédicat Prolog signifiant que l'agent A ressent de la joie avec un degré D au sujet de la proposition Phi à l'instant T

$$\begin{aligned} cond\_emotion(A, joy, [Phi], D, T) :- & \; infocus(A, Phi, Deg, T), \\ & \; believe(A, Phi, D1, T), \\ & \; desire(A, Phi, D2), \\ & \; D \; is \; D1 * D2 * Deg. \end{aligned}$$

Les évaluations ont confirmé l'importance de ce degré d'intensité associé à l'émotion, bien qu'une expression numérique soit peu parlante pour les utilisateurs.

4.1.5. *Expression émotionnelle*

À chaque instant un vecteur de toutes les émotions déductibles de la BC de l'agent est généré. Pour simplifier, nous avons fait l'hypothèse que l'agent ne pouvait exprimer qu'une seule émotion à la fois, même si d'autres auteurs permettent à leur agent d'exprimer un mélange de plusieurs émotions par une expression faciale mixte (Ochs *et al.*, 2005). Nous sélectionnons donc l'émotion la plus appropriée pour l'exprimer, en fonction de son intensité et de sa *complexité* (certaines émotions sont composées de deux émotions plus simples) : pour l'instant nous sélectionnons l'objet de l'émotion la plus intense puis l'émotion la plus complexe (donc plus informative) portant sur cet objet. Remarquons que la sélection de l'émotion à exprimer pourrait aussi dépendre de nombreux paramètres dont nous ne tenons pas encore compte : le contexte (inhibition de certaines émotions en présence de certaines personnes, par exemple on ne doit pas s'énerver contre son supérieur hiérarchique), la culture, le caractère de l'agent...

Dans les paragraphes suivants nous nous intéressons aux applications de cet agent.

4.2. *Évaluation*

La première application de PLEIAD est de permettre d'évaluer la crédibilité de notre modèle logique des émotions. C'est ce qui nous intéresse particulièrement dans cet article. En effet, un agent émotionnel doit manifester des expressions émotionnelles réalistes pour ne pas perdre sa crédibilité. C'est pourquoi notre modèle est fondé sur une théorie psychologique reconnue des émotions humaines : la typologie OCC. Cependant il existe de nombreuses autres théories psychologiques des émotions qui divergent souvent sur leurs définitions. Nous avons donc voulu évaluer non seulement les émotions générées par notre modèle BDI, mais aussi la théorie sous-jacente.

Pour cela nous avons fait évaluer par des utilisateurs humains la pertinence des émotions exprimées par notre agent PLEIAD au cours d'un bref scénario. Nous décrivons ici le déroulement des évaluations puis les conclusions que nous en avons tirées.

4.2.1. *Démarche expérimentale*

PLEIAD dispose d'un mode test dans lequel les modifications de la BC sont contraintes par un scénario prédéfini. À chaque étape, l'utilisateur peut choisir entre plusieurs options qui influencent différemment la BC de l'agent et donc ses émotions. L'émotion correspondante est alors calculée et affichée, et un formulaire d'évaluation permet à l'utilisateur de noter la prévisibilité de l'émotion générée (aurait-il ressenti la même émotion dans cette situation ?) ainsi que sa cohérence (comprend-il que quelqu'un ressente cette émotion dans cette situation ou bien est-ce irréaliste ?). L'évaluateur peut aussi indiquer quelle émotion il aurait ressenti dans ce cas, si elle est différente de celle générée, et donner des commentaires.

Nous avons élaboré un scénario interactif faisant intervenir les douze émotions basées sur les événements de la typologie OCC, l'avons soumis à quinze expérimentateurs et avons recueilli leurs évaluations et leurs commentaires.

4.2.2. *Scénario*

Une personne i souhaite être recrutée sur un emploi qui l'intéresse. Elle essaye pour cela d'obtenir un entretien d'embauche ($Des_i\, entretien$) en envoyant son CV à l'entreprise. Le scénario se déroule en sept étapes :

1) La première option permet de choisir si elle considère que son CV est bon ($Bel_c\, boncv$) ou mauvais ($Bel_c\, \neg boncv$), ou si elle n'a pas d'opinion. Une émotion par anticipation est alors générée au sujet de ses chances d'obtenir un entretien ($Expect_c\, entretien$ ou $Expect_c\, \neg entretien$).

2) À l'étape suivante, elle est convoquée à un entretien ($Bel_c\, entretien$). La deuxième émotion concerne son évaluation de cette convocation.

3) Elle s'intéresse ensuite à ses chances de réussir l'entretien. Une option permet à l'utilisateur de choisir si elle est plutôt pessimiste ($Expect_c\, \neg bonentretien$) ou optimiste ($Expect_c\, bonentretien$). La troisième émotion générée concerne la perspective de réussir ou de rater l'entretien.

4) À l'étape suivante, l'utilisateur doit choisir si elle a réussi ($Bel_c\, bonentretien$) ou raté l'entretien ($Bel_c\, \neg bonentretien$). Une émotion de confirmation est alors générée à ce sujet.

5) La candidate change alors de centre d'attention pour s'intéresser à ses chances d'obtenir l'emploi ($Expect_c\, emploi$ ou $Expect_c\, \neg emploi$). Une nouvelle émotion par anticipation est générée.

6) L'utilisateur peut choisir si elle est embauchée ($Bel_c\, emploi$) ou pas ($Bel_c\, \neg emploi$) sur le poste. Une émotion de confirmation est donc générée.

7) La candidate s'intéresse alors à un autre candidat, qui a eu le poste si elle ne l'a pas eu ($Bel_c\, \neg emploi \wedge Bel_c\, aemploi \wedge Bel_c\, Bel_a\, aemploi$), ou qui ne l'a pas eu si elle l'a eu ($Bel_c\, emploi \wedge Bel_c\, \neg aemploi \wedge Bel_c\, Bel_a\, \neg aemploi$). L'utilisateur choisit s'il s'agit d'un ami ($Undes_c\, Bel_a\, \neg aemploi$), d'un ennemi ($Des_c\, Bel_a\, \neg aemploi$), ou d'un inconnu. Une émotion relative au destin d'autrui est alors générée.

4.2.3. *Résultats*

Les émotions générées sont globalement bien perçues par les utilisateurs, l'agent leur apparaît en général crédible dans la plupart de ses réponses, même si certaines émotions leur sont apparues peu pertinentes, voire aberrantes. Nous n'avons pas recueilli assez d'évaluations pour dresser des statistiques significatives, mais les commentaires nous ont permis de mettre en évidence divers problèmes, que ce soit dans notre formalisme, dans notre interface, ou parfois dans la théorie OCC.

4.2.3.1. La persistance des émotions est mal perçue

Quand le stimulus reçu (le dernier événement produit) ne déclenche aucune nouvelle émotion, nous n'exprimons pas la neutralité en rapport avec ce stimulus. En effet notre module d'expression ne choisit pas l'émotion la plus récente (correspondant au dernier stimulus reçu) mais l'émotion la plus intense. De plus les émotions de l'agent persistent un certain temps après leur déclenchement. Ainsi quand aucune nouvelle émotion n'est déclenchée par un stimulus, l'agent exprime une ancienne émotion persistante. Il semble ainsi répondre à un stimulus par une émotion qui n'y est pas reliée, alors qu'en fait il ne répond pas à ce stimulus, mais continue à exprimer la même émotion que s'il ne s'était pas produit.

Dans une première version de nos travaux (Adam *et al.*, 2005) l'agent exprimait toujours l'émotion déclenchée par le dernier stimulus perçu, mais cela conduisait à des changements trop brutaux de son émotion à chaque nouvelle perception. Ce problème a déjà été mis en évidence par (Moffat *et al.*, 1993). Nous avions donc considéré qu'une solution était de n'exprimer la nouvelle émotion que si son intensité dépassait celle de l'émotion déjà déclenchée. Cependant cette méthode semble peu réaliste au vu des évaluations. Il faudra sans doute envisager d'exprimer un mélange de plusieurs émotions, ou une séquence émotionnelle : d'abord l'émotion reliée au stimulus, puis un retour à l'émotion courante si elle est plus intense.

4.2.3.2. Il manque des émotions dans le modèle OCC

Plusieurs testeurs ont indiqué ressentir plutôt une émotion de surprise dans certaines situations où la typologie OCC génère une émotion de non-confirmation (soulagement ou déception). La surprise n'est pas présente dans la typologie OCC (car non valencée), alors que notamment beaucoup de modèles catégoriels des émotions la placent dans les émotions de base (*e.g.* (Ekman, 1992)). Il faudrait donc apparemment rajouter cette émotion de surprise[13], qui émerge dans des situations non prévues, indépendamment de leur désirabilité. Ainsi dans les cas où l'événement qui vient de se produire était attendu à une probabilité très faible, l'agent sera à la fois surpris et soit soulagé soit déçu. Si l'événement était totalement inattendu, l'agent sera surpris et soit joyeux soit triste. Si l'événement était attendu avec une probabilité suffisante, l'agent sera juste soulagé ou déçu. Cette constatation renforce l'importance de degrés attachés aux différentes attitudes mentales manipulées (ici la probabilité d'un événement).

4.2.3.3. Perception des émotions concernant le destin d'autrui

Dans d'autres cas les testeurs ont même pensé que l'émotion générée par notre formalisation de la typologie OCC était erronée. Dans ce paragraphe et le suivant nous cherchons à comprendre pourquoi l'émotion ressentie par les utilisateurs diffère de celle prédite par notre système : le problème vient-il de notre formalisation de la situation ou de la typologie OCC elle-même ? Comment corriger cette différence ?

13. Ceci pourrait être fait par exemple par la définition suivante : $Surprise_i \varphi \stackrel{déf}{=} Bel_i \varphi \wedge Bel_i P\neg Prob_i \varphi$.

Ainsi, lorsque l'employée apprend qu'un candidat qu'elle déteste (un « ennemi ») a eu l'emploi qu'elle n'a pas eu, nous déclenchons une émotion de ressentiment, définie par l'évaluation d'un événement agréable pour autrui et désagréable pour soi (ici l'embauche du candidat détesté). Or dans cette situation, plusieurs évaluateurs ont déclaré s'attendre plutôt à de la colère de la part de la candidate, alors que selon OCC cette émotion est un mélange de tristesse (ici c'est bien le cas) et de reproche. Le reproche étant déclenché selon OCC par la désapprobation d'une action contraire aux normes, ne peut vraisemblablement pas être déclenché dans cette situation, l'autre candidat ayant le droit de postuler et d'obtenir l'emploi. La typologie OCC semble donc échouer à expliquer la réaction des humains dans cette situation.

Lazarus propose un raffinement de l'émotion de ressentiment définie par OCC : dans le cas où autrui (ami ou ennemi) a obtenu quelque chose que lui désirait aussi mais n'a pas eu, l'individu peut ressentir soit de l'envie, soit de la jalousie si cette ressource convoitée est mutuellement exclusive. Dans la situation où il y a un seul poste à pourvoir, la ressource est mutuellement exclusive, et la candidate devrait donc ressentir de la jalousie. Cependant cette deuxième émotion n'est toujours pas de la colère, et Lazarus semble donc échouer lui aussi à expliquer cette réaction.

Cette différence entre l'émotion prédite par ces deux théories et l'émotion que les évaluateurs ont dit ressentir s'explique peut-être par la difficulté d'étiqueter ses propres émotions : les évaluateurs pouvaient-ils vraiment se mettre à la place de la candidate et dire quelle émotion ils auraient ressentie ? Les prochaines expérimentations seront conduites avec des psychologues pour répondre au mieux à ces questions.

4.2.3.4. Manque de finesse dans les émotions complexes

Une autre émotion qui n'a pas été bien perçue est la « compassion » (*sorry for*) déclenchée pour la candidate quand elle obtient le poste et qu'un ami ne l'obtient pas. En évaluant l'événement indésirable pour son ami de ne pas obtenir le poste, elle est désolée pour lui. Cependant elle est elle-même impliquée dans cet échec puisque c'est elle qui a obtenu le poste. Certaines personnes auraient donc voulu qu'elle soit gênée face à son amie. Cette « gêne » se rapproche un peu d'un sentiment de honte que la typologie OCC prédirait suite à l'évaluation de son action de postuler pour cet emploi si cette action était considérée comme contraire à un standard. Deux facteurs expliquent que la honte n'ait pas été générée par notre formalisme dans cette situation.

Tout d'abord, nous avons choisi de représenter son embauche comme un événement, pour plusieurs raisons : l'action de postuler n'est pas déterministe, son résultat dépend aussi des actions-candidatures des autres agents, et surtout l'embauche est le résultat de l'action du recruteur plus que de celle de l'agent. Comme le recruteur n'est pas considéré dans les accointances de l'agent, on peut traiter son action comme un événement. En formalisant le problème de cette manière, nous avons empêché des émotions de honte ou de remords de se produire. Il faudra sans doute réfléchir plus profondément à la distinction entre événement et action.

De plus, nous ne considérons a priori pas l'action de postuler pour un emploi comme contraire à un standard global. Un tel standard, qui interdirait à tout agent d'envoyer sa candidature pour un emploi, pourrait faire ressentir à un agent une émotion de honte après avoir été embauché, quelles que soient les circonstances, ce qui ne serait pas du tout cohérent. En l'occurrence, on voudrait que la candidate soit gênée devant son ami d'avoir obtenu ce poste à sa place, mais soit tout de même fière devant d'autres personnes, comme le recruteur, ou sa famille. Une piste pour formaliser cet aspect dynamique des standards est de considérer des idéaux paramétrés par un groupe : ces idéaux valent pour un agent faisant partie de ce groupe, quand il est face à ce groupe. Un exemple classique est celui du soldat qui est par ailleurs croyant : sa religion lui interdit de tuer alors que son métier le lui impose parfois ; quand il est amené à tuer il peut ressentir de la honte dans sa communauté religieuse mais sans doute pas face à ses collègues. Ainsi pour en revenir à notre scénario, on peut supposer que dans un groupe d'amis, un standard implicite recommande de ne pas essayer d'obtenir ce qu'un ami convoite aussi. La considération de ce standard implicite permet d'expliquer l'émotion de honte ressentie par les évaluateurs tout en restant dans le cadre des définitions données par OCC.

Par ailleurs, nous voudrions faire remarquer que la typologie OCC définit le remords comme une combinaison de tristesse et de honte. Or dans ce cas le mélange est plus subtil puisque la candidate elle-même est joyeuse du résultat de son action (elle est embauchée), mais a provoqué la tristesse d'un autre agent (qui lui n'est pas embauché), ce qui la rend triste si c'est un ami. On pourrait donc définir une nouvelle émotion (la gêne décrite par les évaluateurs, non présente telle quelle dans la typologie OCC) comme une combinaison de ces éléments, soit pour j un ami de i :

$$Discomfort_{i,j,i:\alpha}\varphi \stackrel{déf}{=} Shame_i\ i{:}\alpha \wedge Bel_i\ Before_{j:\alpha}\ \neg Bel_i\ F\varphi \wedge Joy_i\ \varphi \wedge Bel_i\ Sadness_j\ \varphi$$

Enfin, signalons que la théorie de Lazarus définit plus finement des émotions de honte et de culpabilité bien distinctes qui pourraient aussi nous aider à formaliser plus précisément l'émotion ressentie par les humains dans cette situation.

4.2.3.5. De l'espoir, de la peur, et des probabilités

Un problème récurrent soulevé par les évaluateurs est la confusion entre espoir et peur. Dans quels cas un agent doit-il craindre φ, et dans quels cas doit-il espérer $\neg\varphi$? La distinction semble tenir aux probabilités. Dans certains cas, certaines personnes considèrent plus réaliste de ressentir de l'espoir (resp. de la peur) quand l'événement désirable (resp. indésirable) est peu probable (comme espérer gagner au Loto, ou craindre de se faire écraser en traversant la route ; il paraîtrait bizarre d'avoir peur de perdre, ou d'espérer arriver de l'autre côté de la route). C'est le parti pris par Lazarus dans sa définition de l'espoir. Cependant Ortony, Clore, et Collins définissent l'espoir comme la perspective d'un événement désirable, sa probabilité jouant juste sur l'intensité de l'espoir. Nous avons traduit « perspective » par l'opérateur $Expect$ qui signifie que l'événement est considéré comme probable, et cette définition est donc à l'opposé de celle de Lazarus (que nous avions d'ailleurs adoptée dans une pre-

mière version de nos travaux (Adam *et al.*, 2006a; Adam *et al.*, 2006b)). Ce désaccord montre bien la difficulté de distinguer entre ces deux émotions. Une solution possible serait de déclencher à la fois l'espoir et la peur avec des degrés complémentaires.

Une autre remarque au sujet de ces deux émotions a été qu'elles ne devraient se produire ni quand l'événement était trop probable, ni quand il l'était trop peu. Cette remarque est en accord avec les théories psychologiques que nous utilisons. Cependant il semblerait que dans l'implémentation nous avions mal fixé les seuils de probabilité au-dessus et en-dessous desquels l'émotion ne doit pas apparaître, conduisant parfois à des émotions évaluées comme aberrantes.

4.2.3.6. Au sujet de l'interface

Ces évaluations nous ont aussi permis de mettre en évidence divers problèmes d'ergonomie de l'interface. En particulier l'objet de l'émotion est mal indiqué et n'est souvent pas lu par les évaluateurs. Ainsi ceux-ci considèrent parfois comme aberrante une émotion en se référant à un certain objet alors que l'émotion est déclenchée par rapport à un autre événement, ce qui peut fausser certains résultats.

De plus l'intensité est pour l'instant donnée de manière numérique, par un degré entre 0 et 1, mais cela ne semble pas approprié. Ce chiffre n'est la plupart du temps pas lu par les évaluateurs. Nous envisageons pour la prochaine campagne de tests d'indiquer l'intensité de manière textuelle grâce à trois adverbes : « un peu », « assez », « très » précédant les émotions. Nous simplifierons aussi l'interface et donnerons les informations nécessaires sous une forme plus compréhensible notamment pour des utilisateurs ne maîtrisant pas Prolog. Ainsi au lieu d'afficher « emotion(employee,fear,not(entretien),0.8,0) » nous écrirons plutôt « l'employée est très inquiète à propos de la perspective de ne pas obtenir d'entretien ».

Ces simplifications nous permettront de conduire de nouvelles expérimentations auprès d'utilisateurs non informaticiens. Nous avons l'intention à présent de travailler avec des psychologues pour mener de nouvelles expérimentations plus poussées, notamment avec un plus grand nombre de scénarios disponibles. Cependant cette première évaluation nous a déjà permis d'obtenir des résultats encourageants. Les émotions exprimées apparaissent comme globalement pertinentes, et les commentaires des évaluateurs nous ont permis d'identifier pourquoi certaines émotions n'étaient pas acceptées, soit à cause de leur définition par Ortony *et al.*, soit à cause de la formalisation que nous en avions faite.

4.3. *Intégration dans un ACA et applications*

Une autre application intéressante d'un tel modèle des émotions concerne les agents conversationnels animés (ACA). Nous pensons que le module de génération d'émotions de PLEIAD peut s'intégrer dans n'importe quel ACA à architecture BDI et lui apporter une grande finesse émotionnelle (puisque nous fournissons la formalisation de vingt émotions différentes) rendant ses expressions plus variées.

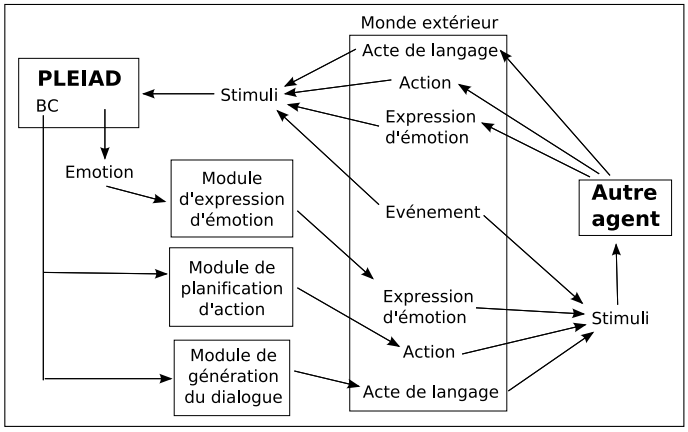

Figure 4. *Intégration de PLEIAD dans un agent BDI*

Nous envisageons divers types d'applications pour un tel agent. Nous travaillons actuellement pour compléter ce modèle de la génération des émotions (processus psychologique d'*appraisal*) par un modèle de leur effet sur la base de connaissances de l'agent (processus de *coping*). Nous envisageons d'inclure ce modèle des émotions dans un modèle du dialogue humain qui serait alors capable de simuler des comportements dialogiques humains dits « irrationnels », rarement pris en compte actuellement notamment à cause de fortes hypothèses de rationalité et de coopération imposées aux agents. Par exemple ces agents conversationnels émotionnels pourraient changer de sujet, couper la parole à leur interlocuteur, refuser de lui répondre, lui mentir, ... De tels agents capables d'adopter un comportement réaliste, humanoïde, influencé par leurs émotions, peuvent intervenir dans des mondes virtuels, que ce soit pour l'entraînement (*e.g*. des pompiers (El Jed *et al.*, 2004) ou des militaires (Gratch *et al.*, 2004)) ou le divertissement (jeux vidéos), afin de favoriser l'immersion de l'utilisateur.

De plus, le modèle BDI des émotions permet à l'agent de se mettre à la place d'un utilisateur humain dont il a un modèle (une description de ses attitudes mentales supposées) pour déduire les émotions que cet utilisateur ressent. Cela a déjà été fait pour des agents pédagogiques (Jaques *et al.*, 2004b; Jaques *et al.*, 2004a) capable d'adapter la stratégie la mieux adaptée à l'émotion de l'apprenant. Nous avons appliqué cette méthode à l'intelligence ambiante (Adam *et al.*, 2006a) pour permettre à un agent intelligent gérant une maison de détecter les émotions de l'habitant et de s'y adapter.

Dans la prochaine partie nous comparons notre modèle théorique à d'autres travaux existants sur la formalisation des émotions.

5. Discussion

Meyer

Meyer (Meyer, 2004) propose un modèle logique des émotions basé sur KARO, sa logique de l'action, de la croyance et du choix. Il utilise cette logique pour écrire des règles de génération pour quatre émotions : la joie, la tristesse, la colère et la peur. Ces conditions dépendent uniquement de la satisfaction des plans de l'agent. En effet le but de Meyer, comme il le dit lui-même, son but n'est pas de capturer les descriptions psychologiques exactement (ou le plus exactement possible) mais plutôt de décrire ce qui fait sens pour des agents artificiels.[14] Au contraire dans nos travaux nous essayons de rester aussi proches que possible des définitions psychologiques des émotions que nous considérons, en nous appuyant fortement sur l'une des approches les plus utilisées en informatique, *i.e.* celle d'Ortony, Clore, et Collins (OCC). Nous fournissons ainsi un formalisme émotionnel plus riche (définissant vingt émotions) et plus fidèle à la psychologie. Cependant ce formalisme est encore limité au déclenchement des émotions, alors que Meyer s'est déjà intéressé à l'influence que les émotions exercent en retour sur les plans de l'agent (Dastani *et al.*, 2006).

Gratch et Marsella

Pour leur agent émotionnel EMA, Gratch et Marsella définissent leur propre structure complexe d'état mental, qu'ils appellent Interprétation Causale. Il s'agit d'une structure en trois parties causalement liées entre elles : l'historique causal (le passé), le monde courant (le présent), et le réseau de tâches (le futur). Les auteurs ont enrichi une représentation classique de planification avec des concepts de théorie de la décision comme la probabilité ou l'utilité, dans le but d'unifier dans une seule architecture tous les besoins d'un agent émotionnel. En effet ils pensent qu'aucun des formalismes précités n'est assez riche à lui seul pour exprimer les émotions d'un agent. Cependant dans nos travaux nous avons montré qu'une logique de type BDI étendue avec des opérateurs de temps, d'action, de probabilité et d'obligation permet de formaliser un grand nombre d'émotions, et nous avons formalisé les deux branches principales de la typologie OCC dans notre logique, et exprimé vingt émotions différentes. Nous espérons ainsi fournir un modèle plus générique et réutilisable des émotions, de nombreux agents étant construits avec une architecture BDI. Le processus d'*appraisal* de Gratch et Marsella, fondé sur la théorie de Lazarus, analyse la configuration de l'Interprétation Causale pour déclencher une ou plusieurs émotions.

La plus intense peut donner lieu à un processus de *coping*, tentative de l'agent de s'adapter à ses émotions. Nous n'avons pas encore formalisé ce processus dans notre logique, mais nous pensons que c'est possible, et des travaux sont en cours dans ce but (Adam *et al.*, 2006d).

14. « Instead of trying to capture the informal psychological descriptions exactly (or as exact as possible), we primarily look here at a description that makes sense for artificial agents. », (Meyer, 2004, p.11)

Ochs *et al*.

Ochs *et al*. (Ochs *et al.*, 2005) s'intéressent plus particulièrement à l'expression faciale d'émotions par des agents animés. Ils donnent une formalisation fondée sur la théorie de l'interaction rationnelle de Sadek (Sadek, 1991; Sadek, 1992), une logique de la croyance, de l'intention et de l'incertitude. Ils essayent de rester très proches de la typologie OCC mais ne décrivent que quatre émotions (joie, tristesse, espoir et peur). Le manque d'une notion de temps empêche de formaliser les émotions de confirmation (déception, soulagement...). Ces travaux ont été enrichis depuis pour rendre compte des émotions empathiques comme « content pour », « désolé pour » (Ochs *et al.*, 2006). Cependant cette approche reste moins complète que la nôtre puisqu'elle ne fournit pas de formalisation des émotions reliées aux actions d'agents. De plus la logique de Sadek inclut une notion de but mais aucune notion de désir ou de préférence personnelle de l'agent, à notre avis indispensable pour exprimer les émotions. Cependant leur formalisme émotionnel intègre la notion de mélange de plusieurs émotions lors de leur expression, alors que notre modèle fait l'hypothèse qu'une seule émotion est exprimée à la fois, hypothèse qui a été contredite par les évaluations comme nous l'avons vu dans la partie précédente.

Jaques *et al*.

Jaques *et al*. (Jaques *et al.*, 2004a; Jaques *et al.*, 2004b) ont développé un agent pédagogique, le *Mediating Agent*, capable d'inférer les émotions d'un étudiant pour adopter une stratégie adaptée et permettre l'apprentissage dans les meilleures conditions. Les auteurs fournissent une formalisation dans le formalisme X-BDI de sept émotions de la typologie OCC (déception, satisfaction, joie, tristesse, gratitude, colère, honte) et une implémentation en Prolog de leur agent. Les règles de déclenchement des émotions se basent sur une notion de désirabilité qui est dépendante du domaine. En effet, les événements possibles dans l'environnement éducatif sont listés et classifiés par rapport à leur désirabilité pour l'étudiant. Cette désirabilité est fixée a priori en fonction du profil de l'étudiant, établi grâce à un questionnaire auquel il répond avant d'utiliser le système. Au contraire, nous proposons une formalisation indépendante du domaine, qui ne repose sur aucune hypothèse particulière ; n'importe quel désir de notre agent est susceptible de déclencher une émotion. Jaques *et al*. affectent ensuite une intensité à l'émotion parmi deux valeurs possibles (moyenne ou forte) en fonction des différents paramètres spécifiés dans la théorie OCC.

6. Conclusion

Nous avons présenté dans cet article un modèle des émotions abouti : nous sommes partis d'une théorie psychologique reconnue, nous l'avons traduite dans un formalisme logique générique et réutilisable tout en restant aussi fidèles que possible à la psychologie, nous avons ensuite implémenté ce modèle théorique dans un agent logiciel, puis nous avons évalué les émotions de cet agent auprès d'humains pour en tirer des conclusions sur la théorie OCC et sur notre modèle BDI.

Les résultats de cette première évaluation nous ouvrent de nombreuses perspectives d'amélioration, du moins pour la partie que nous pouvons modifier, c'est-à-dire notre formalisation de la typologie OCC. Mais par ailleurs, nous souhaitons aussi nous intéresser maintenant à la formalisation dans la même logique d'autres théories psychologiques, afin de comparer leurs prédictions.

En particulier, nous nous intéressons à la théorie de l'*appraisal* de Lazarus, qui nous a souvent semblé plus fine voire plus en accord avec les avis des juges lors des évaluations. Cette théorie est cependant nettement plus complexe que la typologie OCC, n'étant pas créée pour permettre aux chercheurs en IA de l'implémenter. Elle fait notamment intervenir des concepts complexes de responsabilité, d'implication de soi... qui seront difficiles à formaliser dans notre logique BDI actuelle, à moins d'intégrer à notre théorie de l'action le concept de « l'agentitude » (*agency* en Anglais) comme le *seeing-to-it-that* (opérateur STIT) de (Horty *et al.*, 1995) que nous avons par ailleurs commencé à étudier (Herzig *et al.*, 2006; Broersen *et al.*, 2006). Il convient donc de se poser la question du rapport entre les bénéfices obtenus et les coûts supplémentaires engendrés par l'utilisation d'une telle théorie. Est-il nécessaire pour un agent logiciel d'exprimer les fines distinctions entre honte et culpabilité, ou entre jalousie et envie, que Lazarus met en évidence ? Finalement la solution pour ces agents sera peut-être d'utiliser un compromis entre toutes les théories, utilisant parfois des notions simples mais suffisantes, et parfois des définitions plus compliquées sur certaines émotions critiques.

Mais dans tous les cas, nous pensons que dans la mesure où l'on souhaite actuellement, pour diverses raisons, rendre les agents le plus crédibles possible, il est sans doute intéressant d'explorer des théories psychologiques plus complexes que la traditionnelle typologie OCC. De plus la psychologie pourrait peut-être elle-même tirer profit de telles recherches (*cf.* section 4.2.3.2), notamment au travers des possibilités d'évaluation des théories, pour mieux comprendre les émotions humaines.

7. Bibliographie

Adam C., « PLEIAD : ProLog Emotionally Intelligent Agents Designer. Un module de gestion des émotions d'un ACA », *in* J.-C. Martin, C. Pélachaud (eds), *Workshop Francophone sur les Agents Conversationels Animés (WACA)*, Toulouse, 26-27 october, 2006.

Adam C., Evrard F., « Donner des émotions aux agents conversationnels », *in* S. Pesty, J.-P. Sansonnet (eds), *WACA'01 - Premier Workshop francophone sur les Agents Conversationnels Animés*, Grenoble, 13-14 june, 2005.

Adam C., Évrard F., Gaudou B., Herzig A., Longin D., « Modélisation logique d'agents rationnels pour l'intelligence ambiante », *Actes des 14e Journées Francophones sur les Systèmes Multi-Agents (JFSMA 2006), Annecy, France, 18–20 octobre*, vol. à paraître, Hermès Science Publications, 2006a.

Adam C., Gaudou B., Herzig A., Longin D., « A logical framework for an emotionally aware intelligent environment », *in* J. C. Augusto, D. Shapiro (eds), *1st ECAI Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI'06), Riva de Garda, Italy, August*

29th, IOS Press, ftp://ftp.irit.fr/IRIT/LILAC/Adam_aitami2006.pdf, 2006b.

Adam C., Gaudou B., Herzig A., Longin D., « OCC's emotions: a formalization in a BDI logic », *in* J. Euzenat (ed.), *Proc. of the Twelfth Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA'06), Varna, Bulgaria, september 13–15*, vol. 4183 of *LNAI*, Springer-Verlag, p. 24-32, 2006c.

Adam C., Longin D., Coping strategies in a BDI logic: how can agents do to drop their negative emotions, Technical report, IRIT, 2006d. Disponible sur : ftp://ftp.irit.fr/IRIT/LILAC/adam_et_al-TR-2006.pdf.

Anderson J., *The Adaptive Character of Thought*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1990.

Arnold M., *Emotion and personality*, Columbia University Press, New York, 1960.

Becker C., Kopp S., Wachsmuth I., « Simulating the emotion dynamics of a multimodal conversational agent », *ADS'04*, Springer LNCS, 2004.

Berger A., Pesty S., « Vers un langage de conversation entre agents pour l'interaction dans les communautés mixtes », *Actes du Colloque Jeunes Chercheurs en Sciences Cognitives*, Bordeaux, France, 2005.

Broersen J., Herzig A., Troquard N., « A STIT-extension of ATL, with applications in the epistemic and deontic domains », *in* M. Fisher, W. van der Hoek (eds), *Proc. 10th Eur. Conf. on Logics in Artificial Intelligence (JELIA06), Liverpool, 13-15 September 2006*, vol. 4160 of *LNAI*, Springer, 2006.

Burgess J. P., « Basic Tense Logic », *in* D. Gabbay, F. Guenthner (eds), *Handbook of Philosophical Logic*, 2nd edn, vol. 7, Kluwer Academic Publishers, p. 1-42, 2002.

Castelfranchi C., Lorini E., « Cognitive Anatomy and Functions of Expectations. », *IJCAI03 Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions*, Acapulco, Mexico, August 9-11, 2003.

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

Cohen P. R., Levesque H. J., « Intention is Choice with Commitment », *Artificial Intelligence Journal*, vol. 42, n° 2–3, p. 213-261, 1990.

Darwin C. R., *The expression of emotions in man and animals*, Murray, London, 1872.

Dastani M., Meyer J.-J., « Programming Agents with Emotions », *Proc. 17th European Conf. on Artificial Intelligence (ECAI 2006), Trento, Italy, Aug. 28th–Sep. 1st*, IOS Press, 2006.

de Rosis F., Pelachaud C., Poggi I., Carofiglio V., Carolis B. D., « From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent », *International Journal of Human-Computer Studies*, vol. 59, n° 1-2, p. 81-118, 2003.

Ekman P., « An Argument for Basic Emotions », *Cognition and Emotion*, vol. 6, p. 169-200, 1992.

Ekman P., Friesen W., Hager J., *Facial Action Coding System Invetigator's Guide*, A Human Face, 2002.

El Jed M., Pallamin N., Dugdale J., Pavard B., « Modelling character emotion in an interactive virtual environment. », *AISB 2004 Convention: Motion, Emotion and Cognition*, The society for the study of Artificial Intelligence and the Simulation of Behaviour, April, 2004.

Elliott C., The Affective Reasoner : A process model of emotions in a multi-agent system, PhD thesis, Northwestern University, Illinois, 1992.

FIPA, « FIPA Communicative Act Library Specification », , `http://www.fipa.org/repository/aclspecs.html`, 2002. (Foundation for Intelligent Physical Agents).

Forgas J., « Mood and judgment: The affect infusion model (AIM) », *Psychological Bulletin*, vol. 117, p. 39-66, 1995.

Frijda N. H., *The Emotions*, Cambridge University Press, 1986.

Frijda N. H., Ortony A., Sonnemans J., Clore G., « The complexity of intensity: Issues concerning the structure of emotion intensity. », *Review of personality and social psychology*, vol. 11, p. 60-89, 1992.

Gratch J., Marsella S., « A domain independent framework for modeling emotion », *Journal of Cognitive Systems Research*, vol. 5, n° 4, p. 269-306, 2004.

Gratch J., Marsella S., « Lessons from Emotion Psychology for the Design of Lifelike Characters », *Journal of Applied Artificial Intelligence (special issue on Educational Agents - Beyond Virtual Tutors)*, vol. 19, n° 3-4, p. 215-233, 2005.

Grizard A., Lisetti C., « Generation of facial emotional expressions based on psychological theory », *Workshop on Emotion and Computing, KI'2006*, Bremen, Germany, June, 14-19, 2006.

Herzig A., « Modal probability, belief, and actions », *Fundamenta Informaticæ*, vol. 57, n° 2-4, p. 323-344, 2003.

Herzig A., Longin D., « C&L intention revisited », *in* D. Dubois, C. Welty, M.-A. Williams (eds), *Proc. 9th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2004), Whistler, Canada, June 2–5*, AAAI Press, p. 527-535, 2004.

Herzig A., Troquard N., « Knowing How to Play: Uniform Choices in Logics of Agency », *in* G. Weiss, P. Stone (eds), *5th International Joint Conference on Autonomous Agents & Multi Agent Systems (AAMAS-06), Hakodate, Japan, 8–12 mai 2006*, ACM Press, p. 209-216, 2006.

Hintikka J., *Knowledge and Belief: An Introduction to the Logic of the Two Notions*, Cornell University Press, Ithaca, 1962.

Horty J. F., Belnap N., « The deliberate Stit: A study of action, omission, ability, and obligation », *Journal of Philosophical Logic*, vol. 24, n° 6, p. 573-582, 1995.

Jaques P. A., Vicari R. M., « A BDI Approach to Infer Students Emotions », *IBERAMIA*, p. 901-911, 2004a.

Jaques P. A., Vicari R. M., Pesty S., Bonneville J.-F., « Applying Affective Tactics for a Better Learning. », *In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*, IOS Press, 2004b.

Laverny N., Lang J., « From knowledge-based programs to graded belief-based programs, Part II: off-line reasoning », *Proc. of the 9th International Joint Conference on Artificiel Intelligence (IJCAI'05) , Edinburgh, Scotland, 31/07/05-05/08/05*, Gallus, p. 497-502, juillet, 2005.

Lazarus R. S., *Emotion and Adaptation*, Oxford University Press, 1991.

Lebiere C., Anderson J. R., « A Connectionist Implementation of the ACT-R Production System. », *Fifteenth Annual Conference of the Cognitive Science Society*, p. 635-640, 1993.

Meyer J. J., « Reasoning about Emotional Agents », *in* R. L. de Mántaras, L. Saitta (eds), *16th European Conf. on Artif. Intell. (ECAI)*, p. 129-133, 2004.

Moffat D., Frijda N. H., Phaf R. H., « Analysis of a Model of Emotions », *in* A. Sloman, D. Hogg, G. Humphreys, A. Ramsay, D. Partridge (eds), *Prospects for Artificial Intelligence: Proc. of AISB-93*, IOS Press, Amsterdam, p. 219-228, 1993.

Ochs M., Niewiadomski R., Pelachaud C., Sadek D., « Intelligent Expressions of Emotions », *1st International Conference on Affective Computing and Intelligent Interaction ACII*, China, October, 2005.

Ochs M., Pélachaud C., Sadek D., « Les conditions de déclenchement des émotions d'un agent conversationnel empathique », *WACA'2006*, 2006.

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

Picard R. W., *Affective Computing*, MIT Press, 1997.

Reilly N., Believable social and emotional agents, PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, 1996.

Russell J. A., « How shall an emotion be called? », *in* R. Plutchik, H. Conte (eds), *Circumplex models of personality and emotions*, American Psychological Association, Washington, DC, p. 205-220, 1997.

Sadek M. D., Attitudes mentales et interaction rationnelle : vers une théorie formelle de la communication, PhD thesis, Université de Rennes I, Rennes, France, 1991.

Sadek M. D., « A Study in the Logic of Intention », *in* B. Nebel, C. Rich, W. Swartout (eds), *Proc. Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, Morgan Kaufmann Publishers, p. 462-473, 1992.

Sadek M. D., « Dialogue Acts are Rational Plans », *in* M. Taylor, F. Néel, D. Bouwhuis (eds), *The structure of multimodal dialogue*, John Benjamins publishing company, Philadelphia/Amsterdam, p. 167-188, 2000. From ESCA/ETRW, Workshop on The Structure of Multimodal Dialogue (Venaco II), 1991.

Scherer K., « Toward a dynamic theory of emotion: the component process model of affective states », *Geneva studies in Emotion and Communication*, vol. 1, n° 1, p. 1-98, 1987.

Staller A., Petta P., « Introducing emotions into the computational study of social norms: a first evaluation », *Journal of artificial societies and social simulation*, 2001.

Walley P., Fine T. L., « Varieties of modal (classificatory) and comparative probability », *Synthese*, 1979.

Wooldridge M., *Reasoning about rational agents*, MIT Press, 2000.