

Vers un modèle formel des émotions d'un agent rationnel dialoguant empathique

M. Ochs*[†] D. Sadek * C. Pelachaud[†]

*Orange Labs, France
{magalie.ochs, david.sadek}@orange-ftgroup.com

[†]Laboratoire LINC, Université Paris 8, France
c.pelachaud@iut.univ-paris8.fr

Résumé :

Les travaux présentés dans cet article visent à concevoir et mettre en oeuvre des agents rationnels dialoguants capables d'exprimer des émotions, et plus particulièrement des émotions d'empathie, durant leur interaction avec l'utilisateur afin d'améliorer l'interaction humain-machine. Pour ce faire, les agents rationnels dialoguants doivent être capables d'identifier les situations d'interaction dans lesquelles leur interlocuteur peut ressentir des émotions. A partir de la littérature en psychologie cognitive et d'une analyse d'un corpus de dialogues réels humain-machine, nous avons identifié certaines circonstances de déclenchement d'émotions positives et négatives pouvant apparaître dans une interaction humain-machine. Sur cette base, un modèle formel d'émotions d'un agent rationnel dialoguant a été construit. Les conditions de déclenchement d'émotions sont représentées par des états mentaux particuliers, *i.e.* par des combinaisons particulières de croyances, d'incertitudes et d'intentions. L'intensité de l'émotion est calculée à partir de l'état mental de l'agent. Cette formalisation des émotions permet de représenter les émotions d'empathie envers d'autres agents.

Mots-clés : Émotions, empathie, agent rationnel dialoguant

Abstract:

The work presented in this paper aims to develop rational dialog agents able to express emotions, and more particularly empathic emotions, during their interaction with a user in order to enhance human-machine interaction. An empathic rational dialog agent should know the circumstances under which a user may feel an emotion. Relying on psychological theory of emotion elicitation and on a study of real human-machine dialogs during which the user expresses emotions, we have highlighted some situations that may lead to a user's emotion elicitation. From the descriptions of these emotional situations, a formal model of emotions for a rational dialog agent has been designed. The conditions of emotion elicitation are represented in terms of par-

ticular mental states, *i.e.* particular combinations of beliefs, uncertainties, and intentions. The intensity of emotions is computed from the agent's mental state. This formalization of emotions is used to represent empathic emotions.

Keywords: Emotions, empathy, rational dialog agent

1 Introduction

Ces dernières années, un intérêt grandissant est apparu pour la conception et le développement d'*agents conversationnels* capables de dialoguer naturellement avec un utilisateur. Ces agents sont souvent utilisés pour interpréter des rôles typiquement incarnés par des humains, comme par exemple le rôle de *tuteur* [Johnson et al., 2000]. Des recherches récentes ont montré que les expressions d'émotions d'agents virtuels permettent de créer une illusion de vie et ainsi d'augmenter leur crédibilité (traduction du terme anglais *believability*) [Bates, 1994]. De plus, comme l'a souligné Picard [Picard, 1997], l'utilisateur ressent de nombreuses émotions durant son interaction avec un ordinateur. Il peut, par exemple, ressentir et exprimer des émotions négatives lors de défaillances du système informatique ou des émotions positives lorsqu'il réalise une tâche avec succès. Les premières recherches semblent montrer que l'expression d'émotions *empathiques* d'un agent conversationnel permet d'améliorer la perception qu'a l'utilisateur de l'agent, d'induire des

émotions positives et d'augmenter les performances et l'engagement de l'utilisateur dans la réalisation d'une tâche durant l'interaction [Brave et al., 2005, Helmut and Mitsuru, 2005].

Les travaux présentés dans cet article visent à concevoir et mettre en oeuvre des agents conversationnels capables d'exprimer des émotions, et plus particulièrement des émotions d'*empathie*, durant leur communication avec l'utilisateur afin d'améliorer l'interaction humain-machine. L'*empathie* se définit comme "la capacité de se mettre mentalement à la place d'autrui afin de comprendre ce qu'il éprouve" [Pacherie, 2004]. Lors du processus d'*empathie*, un individu simule mentalement une situation vécue par une autre personne ; il s'imagine à sa place (c'est-à-dire avec les mêmes buts et les mêmes croyances et dans cette même situation) et imagine alors l'émotion ressentie par cette personne. Par cette simulation émotionnelle, l'individu peut être amené à ressentir une émotion similaire, appelée dans ce cas *émotion empathique* [Poggi, 2004].

Dans une interaction humain-machine, un agent conversationnel exprime une émotion empathique lorsqu'il pense que dans la situation de l'utilisateur il "ressentirait" la même émotion. Cette croyance sur l'état émotionnel potentiel de l'utilisateur doit être issue non pas de la perception de l'émotion (dans ce cas il s'agirait d'une contagion émotionnelle [Poggi, 2004]), mais de la simulation du processus de déclenchement des émotions de l'utilisateur. En d'autres termes, un agent conversationnel empathique doit, en adoptant la perspective de l'utilisateur, en déduire ses émotions. Par conséquent, il doit connaître les conditions dans lesquelles un individu peut potentiellement ressentir une émotion. À partir d'une analyse de corpus de dialogues humain-machine dans lesquels l'utilisateur exprime des émotions et à la lumière de théories de psychologie cognitive, nous

avons mis en évidence les circonstances de déclenchement de certaines émotions de l'utilisateur. Dans cet article, nous proposons une modélisation et une formalisation de ces émotions et de leurs conditions de déclenchement. Les agents conversationnels auxquels nous nous intéressons plus particulièrement sont *les agents rationnels dialoguants*, des agents de type BDI fondés sur une théorie formelle de l'interaction, appelée Théorie de l'Interaction Rationnelle [Sadek, 1991]. Afin de doter ces agents de la capacité d'inférer les émotions potentiellement ressenties par l'utilisateur durant l'interaction, les conditions de déclenchement d'émotions sont décrites à partir des attitudes mentales de croyance, d'incertitude et d'intention d'un agent rationnel. L'intensité des émotions est calculée à partir de l'état mental de l'agent.

Dans une première partie, nous décrivons les caractéristiques des émotions et de leur condition de déclenchement. Dans une seconde partie, après avoir introduit le concept d'agent rationnel dialoguant, nous présentons une formalisation des émotions et de leur intensité.

2 Les émotions dans une interaction humain-machine

Afin de déterminer à quel moment exprimer une émotion *empathique*, l'agent conversationnel doit connaître les situations d'interaction dans lesquelles l'utilisateur pourrait potentiellement ressentir une émotion. Notre approche pour déterminer les conditions de déclenchement des émotions d'un utilisateur durant l'interaction est fondée à la fois sur des théories cognitives des émotions (appelées *théories de l'évaluation cognitive*) et sur une analyse de corpus de dialogues réels humain-machine où l'utilisateur exprime des émotions.

2.1 Les conditions de déclenchement des émotions

Les théories de l'évaluation cognitive.

Les théories de l'évaluation cognitive (*appraisal*) (e.g. [Scherer, 2000, Roseman, 2001, Ortony et al., 1988]) visent à expliquer ce qui conditionne l'émergence d'une émotion particulière pour un individu donné. Selon ces théories, une émotion est issue de l'évaluation subjective d'un évènement [Scherer, 2000]. Un évènement est générateur d'émotion seulement si l'individu pense que cet évènement affecte un de ses buts [Lazarus, 2001]. Un des éléments déterminant dans le déclenchement d'une émotion est *la relation entre l'évènement et le but* de l'individu (i.e. l'impact de l'évènement sur le but). Par exemple, une émotion de peur est déclenchée chez un individu lorsqu'il pense que son but de survie est menacé. Généralement, une émotion positive est générée quand l'évènement facilite ou permet de réaliser un but. Elle est négative lorsque l'évènement entrave la réalisation d'un but. L'interprétation de l'évènement, et par conséquent l'émotion déclenchée, dépendent principalement des *buts* de l'individu et de ses *croyances* (sur l'évènement et ses conséquences). L'implication de ces attitudes mentales propres à chaque individu dans le déclenchement des émotions permet d'expliquer les réactions émotionnelles différentes d'individus distincts face à une même situation.

L'analyse des corpus de dialogues humain-machine.

Afin d'identifier plus concrètement les conditions de déclenchement des émotions d'un utilisateur lors de son interaction avec un agent conversationnel, nous avons analysé des dialogues qui ont amené l'utilisateur à exprimer des émotions. L'objectif est de déterminer les caractéristiques des situations dialogiques génératrices d'émotions dans le contexte de dialogue humain-

machine. Les dialogues ont été annotés afin de mettre en évidence les croyances et les buts des utilisateurs dans les situations émotionnelles. Les dialogues analysés sont issus de deux applications vocales développées à Orange Labs où l'utilisateur interagit oralement avec un agent dialoguant pour obtenir une information dans un domaine particulier (transactions boursières, guide de restaurants). Un schéma de codage, fondé à la fois sur des théories de l'évaluation cognitive [Scherer, 2000, Ortony et al., 1988] et sur la théorie des actes de langage [Austin, 1962], a été utilisé pour l'annotation (pour plus de détails sur le schéma de codage, des exemples de dialogues et l'analyse des corpus voir [Ochs et al., 2006]). L'analyse des dialogues annotés a permis de mettre en évidence les hypothèses ci-dessous sur les situations génératrices d'émotions *negatives*¹ dans une interaction humain-machine. Un évènement peut être déclencheur d'*émotions négatives* chez l'utilisateur lorsqu'il entraîne l'une des situations suivantes :

- *l'échec d'une tentative de satisfaction d'intention*² de l'utilisateur qu'il pensait pouvoir satisfaire ;
- *un conflit de croyance sur une intention* : l'utilisateur pense que l'agent conversationnel considère que l'utilisateur a une intention que celui-ci n'a pas.

Les théories de l'évaluation cognitive et les hypothèses issues de l'analyse de corpus de dialogues permettent de mettre en évidence certaines situations dans lesquelles des émotions (positives ou négatives) d'un utilisateur peuvent être déclenchées. Par ailleurs, un agent conversationnel empathique doit aussi être capable de se représenter une émotion.

¹Dans le corpus de dialogues, aucun cas d'expression d'émotion positive n'a pu être étudié.

²Dans les dialogues humain-machine étudiés, nous avons plus particulièrement observé les *intentions* de l'utilisateur et de l'agent. Une intention est un but persistant d'*avoir agi pour atteindre une situation donnée* [Sadek, 1991].

2.2 La représentation des émotions d'un agent conversationnel

Généralement, une émotion est représentée par différentes caractéristiques. Nous présentons ci-dessous celles nécessaires à la description des émotions d'un agent conversationnel.

Les types d'émotion.

Une émotion est généralement décrite par son *type* (comme par exemple la joie, la satisfaction, la colère, la frustration, etc.). Selon les théories de l'évaluation cognitive [Scherer, 2000, Ortony et al., 1988], ce sont les conditions de déclenchement de l'émotion qui déterminent son *type*. Le type d'une émotion renseigne généralement sur la valence (positive versus négative) de l'émotion. Dans cet article, nous distinguons les types d'émotion suivant leur valence : nous regroupons les types d'émotion positive (*respectivement négative*) sous le terme *émotion positive* (*respectivement négative*).

Les émotions dirigées vers autrui.

Certains types d'émotion ont comme cible autrui. Par exemple, on est en colère contre quelqu'un ou on admire quelqu'un. Ces types d'émotions sont alors caractérisées par la personne vers qui est dirigée l'émotion.

Les *émotions d'empathie*, quelles que soit leur type, sont, elles aussi, par définition des émotions dirigées vers une autre personne, celle *pour* laquelle on a de l'empathie. On est par exemple joyeux *pour* quelqu'un ou triste *pour* quelqu'un d'autre. Elles sont donc caractérisées par la personne vers qui est dirigée l'émotion d'empathie. De plus, dans le cas de certains types d'émotions d'empathie comme la colère ou l'admiration, l'émotion est dirigée vers deux individus distincts : l'individu pour qui on a de l'empathie et l'individu cible. Par exemple, dans le cas de la

colère, un individu a une émotion d'empathie de colère pour un individu *a* contre un individu *b*.

Comme dans le modèle OCC [Ortony et al., 1988], nous distinguons les émotions empathiques des émotions non empathiques. Par conséquent, le fait qu'un agent soit joyeux pour quelqu'un ne signifie pas qu'il a une émotion non empathique de joie.

L'intensité d'une émotion.

A une émotion est généralement associée une valeur numérique représentant son intensité. L'intensité des émotions est déterminée par des valeurs de variables appelées *variables d'intensité* [Ortony et al., 1988]. Dans le contexte du dialogue humain-machine, nous considérons les variables d'intensité suivantes :

- le *degré de certitude* d'une information représente la probabilité qu'une information soit vraie selon l'individu. D'après notre analyse de corpus (§ 2.1), dans le cas de l'échec d'une tentative de satisfaction d'une intention, l'intensité de l'émotion négative semble être *proportionnelle au degré de certitude* : plus un agent était certain (avant l'évènement) de pouvoir satisfaire son intention par l'évènement qui vient d'avoir lieu, plus l'émotion négative générée par l'échec est forte. A l'inverse, nous supposons, fondé sur le modèle OCC [Ortony et al., 1988], que l'intensité d'une émotion positive est *inversement proportionnelle au degré de certitude* : plus un agent était incertain avant l'évènement (*i.e.* plus le degré de certitude était faible) de pouvoir satisfaire son intention par l'évènement qui vient d'avoir lieu, plus l'émotion positive générée par la satisfaction de l'intention est forte.
- l'*effort investi* par un individu pour tenter d'atteindre un but va influencer l'intensité de l'émotion déclenchée. L'intensité de l'émotion est généralement d'autant plus forte que l'effort pour

tenter de satisfaire le but est important [Ortony et al., 1988]. Ainsi, lors de l'échec d'une tentative de satisfaction d'une intention d'un individu, l'émotion déclenchée sera d'autant plus forte qu'il aura investi beaucoup d'effort pour tenter de la satisfaire.

- le *potentiel de réaction* : lors de l'échec d'une tentative de satisfaction d'une intention, nous émettons l'hypothèse que si l'individu pense pouvoir satisfaire son intention par une autre action, l'intensité de l'émotion déclenchée est moins forte.
- *l'importance pour l'individu que son intention soit satisfaite* : lorsqu'un évènement permet la satisfaction ou engendre l'échec d'une tentative de satisfaction d'une intention de l'individu, l'intensité de l'émotion est proportionnelle à l'importance pour l'individu que cette intention soit satisfaite. Typiquement, l'intention de "fermer une porte" est généralement moins importante que celle d'"être heureux". Lors de l'échec de l'intention de "fermer la porte", l'intensité de l'émotion déclenchée est moins forte que dans le cas de l'échec de l'intention d'"être heureux".

En résumé, une émotion peut être représentée par ses *conditions de déclenchement* lesquelles vont déterminer son *type*, sa *direction* et son *intensité*.

3 Modélisation et formalisation des émotions d'un agent rationnel dialoguant

A partir de la description des caractéristiques d'une émotion introduite ci-dessus, un modèle formel de l'émotion fondé sur un modèle des états mentaux d'un agent rationnel dialoguant a été construit. Après une introduction du concept d'agent rationnel dialoguant, nous présentons plus en détails la modélisation et la formalisation des émotions.

3.1 Le concept d'agent rationnel dialoguant

Nous nous appuyons sur un modèle d'agent rationnel fondé sur une théorie formelle de l'interaction (appelée Théorie de l'Interaction Rationnelle [Sadek, 1991]) reposant sur une approche de type BDI. Un agent rationnel dialoguant utilise les attitudes mentales suivantes pour raisonner et agir sur son environnement :

- *La croyance* : une proposition constitue une croyance d'un agent si celui-ci considère que cette proposition est vraie. La croyance est l'attitude mentale par laquelle un agent dispose d'un modèle du monde dans lequel il évolue.
- *L'incertitude* : une proposition constitue une incertitude d'un agent si celui-ci n'est pas tout à fait certain que cette proposition est vraie.
- *Le choix* : une proposition constitue un choix d'un agent si celui-ci préfère que le monde actuel satisfasse cette proposition.
- *L'intention* : une proposition constitue l'intention d'un agent lorsque (1) il pense que la proposition n'est actuellement pas vérifiée, (2) il désire de façon persistante que cette propriété soit réalisée jusqu'à ce qu'il pense cette proposition satisfaite ou impossible à satisfaire et (3) il souhaite accomplir le début de toute séquence d'actions (éventuellement multi-agent) qui peut aboutir à la satisfaction de la proposition.

Dans la Théorie de l'Interaction Rationnelle [Sadek, 1991], les concepts d'attitudes mentales décrits ci-dessus sont formalisés dans le cadre d'une logique modale du premier ordre. Nous introduisons brièvement les aspects du formalisme dont nous nous servons. Dans la suite les symboles \neg , \wedge , \vee , \Rightarrow et \Leftrightarrow représentent les connecteurs logiques classiques de négation, conjonction, disjonction, implication et équivalence. Les symboles \exists et

\forall représentent les quantificateurs existentiels et universels, ϕ et ψ des formules, c, c_1 des variables numériques, i, j et k des variables schématiques dénotant des agents, $type$ une variable représentant un type d'émotion, e, e_1, e_2 des séquences d'évènements éventuellement vides. Les attitudes mentales de croyance, d'incertitude et de choix sont formalisées respectivement par les opérateurs modaux B, U et C tel que $B_i\phi$ peut être lue comme "l'agent i pense que ϕ est vraie"; $U_{i,pr}\phi$ peut être lue comme "l'agent i considère que ϕ a une probabilité pr d'être vraie" avec $pr \in]0, 1[$; $C_i\phi$ peut être lue comme "l'agent i a le désir que ϕ soit vraie". L'opérateur modal composite d'intention I est défini à partir des opérateurs de croyance et de choix. La formule $I_i\phi$ peut être lue comme "l'agent i a l'intention que ϕ soit vraie".

Un agent passe d'un état mental à un autre suite à l'occurrence d'un évènement. La notion de temps est définie par rapport aux évènements et formalisée à travers les opérateurs *Faisable* et *Fait*. *Faisable*(e, ϕ) signifie que l'évènement e peut avoir lieu après quoi ϕ sera vraie. Cette opérateur décrit le futur proche. La formule *Fait*(e, ϕ) signifie que l'évènement e vient juste d'avoir lieu avant quoi ϕ était vraie (*Fait*(e) \equiv *Fait*($e, vrai$)). Cet opérateur décrit le passé proche. La notion de souvenir permet à un agent de comparer ses croyances courantes à ses croyances antérieures à un évènement. Le souvenir de la croyance d'une proposition ϕ d'un agent i avant un évènement e est formalisé par l'attitude mentale de croyance suivante : $B_i(Fait(e, B_i\phi))$. L'abréviation *Unitaire*(e) signifie que e dénote un évènement unitaire. La formule *Agent*(i, e) est vrai si et seulement si l'agent i est l'auteur de l'évènement e .

Les opérateurs $B, C, Faisable$ et *Fait* obéissent à une sémantique des mondes possibles avec pour chaque opérateur une relation d'accessibilité. La logique de la croyance est KD45 (pour plus de détails

voir [Sadek, 1992]).

3.2 Modélisation et formalisation des émotions d'un agent rationnel dialoguant

La représentation d'une émotion déclenchée.

Un évènement ayant lieu dans l'environnement d'un agent peut générer une émotion lorsqu'il affecte une de ses intentions (ou une intention de son interlocuteur)³. Nous appelons **émotion déclenchée** une émotion qui vient d'être déclenchée par un évènement. Elle est représentée par son *type*, son *intensité*, l'*agent chez qui l'émotion a été déclenchée*, l'*agent vers qui elle est dirigée*, l'*évènement* qui l'a déclenché et l'*intention* affectée par l'évènement.

Pour modéliser les *émotions déclenchées non empathiques*, le langage logique est enrichi d'un opérateur modal d'émotion *Emotion_i* pour chaque agent i . La formule *Emotion_i*($type, c, j, e, \phi$) peut être lue comme "l'agent dénoté par i a une émotion non empathique de type $type$ et d'intensité c envers l'agent j ; cette émotion est déclenchée par l'évènement e ayant affecté l'intention de l'agent i de réaliser la propriété dénotée par ϕ ". Lorsque i et j désigne le même agent, la formule représente une émotion *non dirigée* vers un autre agent (§ 2.2). En effet, l'agent dénoté par i représente à la fois celui chez qui l'émotion est déclenchée et celui vers qui elle est dirigée (une émotion non dirigée vers un autre agent est représentée par une émotion dirigée vers l'agent lui-même).

Une émotion déclenchée d'*empathie* de l'agent i pour l'agent j est représentée par l'opérateur modal *Emotion_{emp_{i,j}}*. La formule *Emotion_{emp_{i,j}}*($type, c, k, e, \phi$)

³Dans cet article, nous nous intéressons exclusivement aux émotions déclenchées lorsqu'une intention de l'agent est affectée. Nous ne prenons pas en compte les émotions reliées aux choix (au sens défini dans [Sadek, 1991]) et standards (au sens défini dans [Ortony et al., 1988]) de l'agent

peut être lue comme “l’agent dénoté par i a une émotion d’empathie envers l’agent j de type $type$ et d’intensité c dirigée vers l’agent k , cette émotion est déclenchée par l’évènement e ayant affecté l’intention de réaliser la propriété dénotée par ϕ de l’agent j ”. Le type de l’émotion représentée est *non dirigée* si j et k désignent le même agent. Pour représenter l’émotion d’empathie de l’agent i pour l’agent j dirigée vers l’agent k (comme par exemple l’empathie de l’agent i pour l’agent j de colère contre l’agent k), j et k doivent être distincts.

Une émotion déclenchée se définit par ses conditions de déclenchement, lesquelles vont déterminer le type de l’émotion, l’agent chez qui l’émotion est déclenchée, l’agent vers qui est dirigée l’émotion, l’évènement déclencheur et l’intention affectée. L’intensité de l’émotion dépend de ces paramètres. Nous introduisons ci-dessous une modélisation et formalisation des variables d’intensité utilisées dans la suite pour calculer l’intensité de l’émotion déclenchée. Nous définissons ensuite formellement les émotions déclenchées.

L’intensité des émotions déclenchées.

A partir des descriptions des variables d’intensité (présentées en § 2.2), nous avons modélisé et formalisé ces dernières comme suit.

Le *degré de certitude* de l’agent i concernant la faisabilité d’une proposition ϕ par un évènement e est noté $deg_cert(i, e, \phi) \in [0, 1]$ tel que :

$$deg_cert(i, e, \phi) = \begin{cases} 0 & \text{ssi } B_i(\neg Faisable(e, \phi)) \\ d_c \in]0, 1[& \text{ssi } U_{i,d_c}(Faisable(e, \phi)) \\ 1 & \text{ssi } B_i(Faisable(e, \phi)) \end{cases}$$

En d’autres termes, si un agent pense que la proposition ϕ n’est pas satisfiable par l’évènement alors son degré de certitude est nul. Dans le cas contraire, le degré de certitude est égal à 1. Enfin, si l’agent est

incertain quant à la satisfaction de la proposition ϕ par l’évènement alors le degré de certitude est égal à la probabilité avec laquelle l’agent pense cette proposition faisable.

On note $potentiel_reaction(i, \phi)$ le potentiel de réaction de l’agent i face à l’échec d’une tentative de satisfaction d’une intention ϕ . Pour le calcul du potentiel de réaction, nous proposons les formules suivantes :

$$potentiel_reaction(i, \phi) = \begin{cases} 0 & \text{ssi } B_i(\forall e \neg Faisable(e, \phi)) \\ d_c & \text{ssi } d_c = max \\ & \{proba|U_{i,proba}(Faisable(e, \phi))\} \\ 1 & \text{ssi } (\exists e B_i(Faisable(e, \phi))) \end{cases}$$

En d’autres termes, si un agent pense qu’il n’existe pas d’évènement permettant de satisfaire son intention qui vient d’échouer, le potentiel de réaction est nul. Dans le cas contraire, il est égal à la plus haute probabilité selon l’agent qu’une séquence d’évènements lui permet de satisfaire son intention. Le potentiel de réaction est égal à 1 si l’agent pense qu’il existe une séquence d’évènements permettant de satisfaire son intention.

On définit l’*effort* d’un agent i pour tenter de satisfaire une intention ϕ (noté $effort(i, \phi)$) par le nombre d’actions effectuées par l’agent pour tenter de satisfaire son intention ϕ :

$$effort(i, \phi) = n, n \in N \text{ tel que} \\ \text{Soit } Evt = \{e_1, \dots, e_m\} \text{ tel que} \\ B_i(Fait(e_1; \dots; e_m, \\ Faisable(e_1; \dots; e_m, \phi))) \\ n = card\{e \in Evt, Unitaire(e) \wedge \\ \exists e', e'' Fait(e'; e; e'') \wedge Agent(i, e)\}$$

Remarque : Les séquences d’évènements e' et e'' (pouvant être vides) sont introduites dans la formule ci-dessus afin de caractériser l’ensemble des évènements réalisés par l’agent lui-même et pas uniquement le dernier évènement qui vient d’être

réalisé par l'agent (qui se traduirait par la formule $Fait(e) \wedge Agent(i, e)$).

L'importance d'une intention ϕ pour un agent i notée $imp(i, \phi)$ est un nombre réel positif ($imp(i, \phi) \in \mathbb{R}^+$). Cette valeur représente l'importance pour l'agent que son intention soit satisfaite. Elle doit être fixée par le concepteur mais cela peut dériver d'un modèle de préférences de l'agent.

La fonction d'intensité détermine l'intensité d'une émotion suivant le degré de certitude, le potentiel de réaction, l'effort et l'importance de l'intention. Ces quatre éléments constituent les paramètres de la fonction. Nous proposons la fonction d'intensité $f_intensite$ suivante :

$$\begin{aligned} f_intensite(deg_cert(i, e, \phi), \\ potentiel_reaction(i, \phi), \\ effort(i, \phi), imp(i, \phi)) = \\ deg_cert(i, e, \phi) * potentiel_reaction(i, \phi) * \\ effort(i, \phi) * imp(i, \phi) \end{aligned}$$

Définitions formelles des émotions déclenchées.

Fondées sur la littérature et sur notre analyse d'un corpus de dialogues (§2.1), les conditions de déclenchement des émotions ainsi que leur intensité sont modélisées et formalisées comme suit.

Nous introduisons tout d'abord quelques définitions nous permettant de décrire dans la suite les conditions de déclenchement des émotions :

- l'échec d'une tentative de satisfaction d'une intention ; soit ϕ une intention de l'agent i , e l'évènement qui vient juste d'avoir lieu :

$$\begin{aligned} echec_intention_i(e, \phi) \equiv^{def} \\ B_i(Fait(e, I_i\phi \wedge (U_{i,p,r}(Faisable(e, \phi)) \\ \vee B_i(Faisable(e, \phi)))) \wedge \neg\phi) \end{aligned}$$

L'échec d'une tentative de satisfaction d'une intention signifie ainsi que (1) l'agent i pense qu'un évènement e

vient de se produire ($B_i(Fait(e))$), (2) l'agent avait avant l'évènement e l'intention ϕ ($I_i\phi$), (3) il pensait avec une probabilité p_r (ou il était certain de) pouvoir satisfaire son intention ϕ par l'évènement e ($U_{i,p,r}(Faisable(e, \phi)) \vee B_i(Faisable(e, \phi))$) et (4) après l'occurrence de l'évènement e , l'intention ϕ de l'agent n'est toujours pas satisfaite ($B_i(\neg\phi)$).

- la satisfaction d'une intention ; soit ϕ une intention de l'agent i , e l'évènement qui vient juste de se produire :

$$\begin{aligned} real_intention_i(e, \phi) \equiv^{def} \\ B_i(Fait(e, I_i\phi \wedge \neg B_i(Faisable(e, \phi)) \wedge \phi) \end{aligned}$$

La formule de satisfaction d'une intention signifie que (1) l'agent i pense qu'un évènement e vient d'avoir lieu ($B_i(Fait(e))$), (2) l'agent avait avant l'évènement e l'intention ϕ ($I_i\phi$), (3) il n'avait pas la croyance que l'occurrence de l'évènement e allait permettre la satisfaction de son intention ϕ ($\neg B_i(Faisable(e, \phi))$) et (4) après l'occurrence de l'évènement e , l'intention ϕ de l'agent est satisfaite ($B_i(\phi)$).

- le conflit de croyance sur une intention apparaît lorsqu'un agent considère que son interlocuteur pense qu'il a une intention particulière que l'agent ne pense pas avoir. Soit ϕ une propriété, i un agent et j son interlocuteur, e l'évènement qui vient juste d'avoir lieu :

$$\begin{aligned} conflit_croyance_int_i(e, \phi, j) \equiv^{def} \\ B_i(Fait(e, \neg B_j(I_i(\phi)) \wedge \neg I_i(\phi)) \\ \wedge B_j(I_i(\phi)) \wedge \neg I_i(\phi)) \end{aligned}$$

La formule de conflit de croyance sur une intention signifie que l'agent i pense qu'un évènement e vient de se produire ($B_i(Fait(e))$). Avant cet évènement, l'agent i n'avait pas l'intention ϕ ($\neg I_i(\phi)$) et pensait que l'agent j n'avait pas la croyance qu'il

avait cette intention ($B_i(\neg B_j(I_i(\phi)))$). Après l'évènement e , l'agent i n'a toujours pas l'intention ϕ mais pense que l'agent j croit qu'il a cette intention ($B_i(B_j(I_i(\phi)))$)

Les émotions déclenchées positives non empathiques.

Une *émotion positive non empathique* (et non dirigée) d'intensité c est déclenchée chez l'agent i par un évènement e par rapport à une intention ϕ (notée $Emotion_i(pos, c, i, e, \phi)$) lorsque l'évènement a entraîné la satisfaction d'une intention de l'agent :

$$Emotion_i(pos, c, i, e, \phi) \equiv^{def} real_intention_i(e, \phi)$$

avec

$$c = f_intensite(1 - deg_cert(i, e, \phi), 1, effort(i, \phi), imp(i, \phi))$$

L'intensité de l'émotion est alors proportionnelle à "1 - le degré de certitude de l'agent quant à la faisabilité de son intention par l'évènement", proportionnelle à l'effort investi par l'agent et à l'importance pour lui que l'intention soit satisfaite. Le potentiel de réaction n'est calculé que lors de l'échec d'une intention (§ 2.2) (la valeur passée en paramètre de la fonction est donc 1).

Dans le modèle présenté ici, nous ne définissons pas les émotions déclenchées positives dirigées vers un autre agent (telle que l'admiration par exemple). Ces types d'émotion semblent apparaître rarement dans le cadre d'une interaction humain-machine.

Les émotions déclenchées négatives non empathiques.

Une *émotion négative* (et non dirigée) d'intensité c est déclenchée chez l'agent i par un évènement e par rapport à une intention ϕ (notée $Emotion_i(neg, c, i, e, \phi)$)

si l'évènement a entraîné l'échec d'une tentative de satisfaction d'une intention de l'agent :

$$Emotion_i(neg, c, i, e, \phi) \equiv^{def} echec_intention_i(e, \phi)$$

avec

$$c = f_intensite(deg_cert(i, e, \phi), 1 - potentiel_reaction(i, \phi), effort(i, \phi), imp(i, \phi))$$

L'intensité de l'émotion est proportionnelle au degré de certitude, à l'effort et à l'importance de l'intention et à "1 - le potentiel de réaction".

Une *émotion négative* causée par un autre agent est *dirigée* contre ce dernier. Une émotion négative de l'agent i envers l'agent j d'intensité c déclenchée par un évènement e ayant affecté la satisfaction d'une intention ϕ de l'agent est notée $Emotion_i(type, c, j, e, \phi)$. Cette émotion est déclenchée lorsque l'échec d'une tentative de satisfaction d'intention (une émotion déclenchée négative) est causé par un autre agent suite à *un conflit de croyance sur cette intention*.

$$Emotion_i(neg, c, j, e, \phi) \equiv^{def} conflit_croyance_int_i(e, \psi, j) \wedge Emotion_i(neg, c, i, e, \phi)$$

avec

$$c = f_intensite(deg_cert(i, e, \phi), 1 - potentiel_reaction(i, \phi), effort(j, \psi) + effort(i, \phi), imp(i, \phi))$$

L'intensité de l'émotion est proportionnelle au degré de certitude, à l'importance de satisfaire l'intention, aux efforts de l'agent j et i et à "1 - le potentiel de réaction".

Cette formalisation des émotions permet à un agent rationnel dialoguant d'identifier les situations dans lesquelles une émo-

tion est potentiellement déclenchée chez un autre agent.

Les émotions déclenchées empathiques.

L'agent rationnel dialoguant est *empathique*. Ainsi, le fait que l'agent a une émotion d'empathie envers un autre agent signifie qu'il pense que ce dernier a une émotion particulière. Nous définissons les émotions déclenchées d'empathie comme suit :

$$\begin{aligned} Emotion_emp_{i,j}(type, c, k, e, \phi) &\equiv^{def} \\ B_i(Emotion_j(type, c, k, e, \phi) \wedge \gamma \end{aligned}$$

En d'autres termes, le fait que l'agent i a une émotion d'empathie pour l'agent j de type $type$ dirigée vers l'agent k et d'intensité c suite à l'évènement e ayant affecté une intention ϕ signifie que l'agent i pense que l'agent j a une émotion déclenchée (non empathique) de même type et de même intensité envers l'agent k suite à l'évènement e ayant affecté une intention ϕ de j . Les conditions de déclenchement d'une émotion d'empathie représentée par γ doivent être vérifiées. Elle représente le fait que l'agent i "aime bien" (au sens défini dans [Ortony et al., 1988]) l'agent j . Nous les supposons vraies dans notre modèle.

Remarque : (1) Nous ne nous limitons pas aux deux types d'émotion d'empathie *content pour quelqu'un* et *désolé pour quelqu'un* introduits dans [Ortony et al., 1988]. En effet, une émotion d'empathie est par définition [Poggi, 2004] du type de l'émotion ressentie par celui envers qui l'émotion d'empathie est dirigée. Ainsi, on peut par exemple *avoir peur pour quelqu'un*. (2) Nous supposons que l'émotion d'empathie est de même intensité que l'émotion de l'agent vers qui est dirigée l'émotion d'empathie. Pour affiner le modèle, une fonction pour le calcul de l'intensité de l'émotion d'empathie suivant l'intensité de l'émotion de l'agent vers qui est dirigée cette émotion pourrait être introduite.

Axiomes propres.

Dans le contexte d'une interaction humain-machine, on peut souhaiter que si l'agent pense que son interlocuteur a une émotion positive ou négative non empathique dirigée vers lui alors l'agent lui-même aura cette émotion. Par exemple, si l'agent pense que l'utilisateur est en colère contre lui alors l'agent sera en colère contre lui-même. Ceci se traduit par l'axiome suivant :

$$\begin{aligned} B_i(Emotion_j(type, c, i, e, \phi)) \\ \Rightarrow Emotion_i(type, c, i, e, \phi) \end{aligned}$$

De plus, nous ne souhaitons pas qu'un agent puisse adopter l'intention qu'un autre agent ressent des émotions négatives. Nous imposons donc au modèle l'axiome suivant :

$$\neg I_i(Emotion_j(neg, c, k, e, \phi))$$

Théorèmes.

Étant donné l'axiome ci-dessus, une émotion positive (*resp.* négative) ne peut être déclenchée par une émotion négative (*resp.* positive) d'un autre agent.

$$\begin{aligned} \vdash \neg Emotion_i(pos, c, i, e, \\ (Emotion_j(neg, c_1, z, e_1, \phi))) \end{aligned}$$

$$\begin{aligned} \vdash \neg Emotion_i(neg, c, z, e, \\ (Emotion_j(pos, c_1, j, e_1, \phi))) \end{aligned}$$

Un même évènement ne peut pas déclencher à la fois une émotion positive et négative par rapport à une même intention :

$$\begin{aligned} \vdash \neg (Emotion_i(pos, c, i, e, \phi) \wedge \\ Emotion_i(neg, c_1, j, e, \phi)) \end{aligned}$$

Il en est de même pour les émotions d'empathie :

$$\begin{aligned} \vdash \neg (Emotion_emp_{i,j}(pos, c, j, e, \phi) \wedge \\ Emotion_emp_{i,j}(neg, c_1, z, e, \phi)) \end{aligned}$$

La preuve découle des définitions des émotions empathiques et non empathiques.

L'agent est capable de s'*introspecter* sur ses propres émotions :

$$\vdash Emotion_i(type, c, j, e, \phi) \Leftrightarrow B_i(Emotion_i(type, c, j, e, \phi))$$

$$\vdash \neg Emotion_i(type, c, j, e, \phi) \Leftrightarrow B_i(\neg Emotion_i(type, c, j, e, \phi))$$

La preuve découle des définitions des émotions et du fait que la logique régissant l'opérateur de croyance est de type KD45.

Les modèles d'émotions existants.

Étant donnée l'étroite relation entre le déclenchement d'une émotion et les croyances et les buts d'un individu, des modèles d'émotions construits à partir d'attitudes mentales ont d'ores et déjà été proposés [Dyer, 1987, DeRosis et al., 2003, Meyer, 2006, Adam et al., 2006]. Les travaux présentés dans cet article se distinguent de ces derniers principalement par l'originalité de la formalisation des émotions empathiques, non empathiques et des variables d'intensité. De plus, contrairement aux modèles d'émotions existants, les conditions de déclenchement d'émotions auxquelles nous nous intéressons sont fondées à la fois sur des théories en psychologie cognitive et sur une analyse de corpus de dialogues.

4 Conclusion et perspectives

Pour être capable d'exprimer des émotions d'empathie envers un utilisateur, un agent rationnel dialoguant doit pouvoir identifier les situations d'interaction dans lesquelles son interlocuteur peut ressentir des émotions observables. A partir de la littérature en psychologie cognitive et d'une analyse d'un corpus de dialogues réels

humain-machine, nous avons identifié certaines conditions dans lesquelles des émotions positives et négatives peuvent apparaître. Sur ces bases, un modèle formel d'émotions d'un agent rationnel dialoguant a été construit. Les émotions sont définies par leur condition de déclenchement lesquelles sont représentées par des états mentaux particuliers, *i.e.* par des combinaisons particulières de croyances, d'incertitudes et d'intentions. L'intensité de l'émotion est calculée à partir de cet état mental. Cette formalisation permet la représentation des émotions d'empathie envers d'autres agents. Les conditions de déclenchement d'émotions positives et négatives utilisées peuvent être enrichies afin de formaliser des types d'émotions plus fins, comme la joie, la satisfaction ou la frustration.

Ce modèle d'émotions a été intégré dans un agent rationnel dialoguant couplé avec un visage parlant capable d'adopter différentes expressions faciales suivant l'émotion déclenchée. La prochaine étape vise à évaluer, dans des situations réelles de dialogue d'utilisateurs avec cet agent, la pertinence des conditions de déclenchement des émotions d'empathie de l'agent ainsi que leur impact sur la satisfaction de l'utilisateur et sa perception du système.

Références

[Adam et al., 2006] Adam, C., Gaudou, B., Herzig, A., and Longin, D. (2006). Occ's emotions : a formalization in a bdi logic. In *the Proceedings of the International Conference on Artificial Intelligence : Methodology, Systems, Applications*.

[Austin, 1962] Austin, J. (1962). *How to do things with words*. Oxford University Press, London.

[Bates, 1994] Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM (CACM)*, 37(7) :122–125.

- [Brave et al., 2005] Brave, S., Nass, C., and Hutchinson, K. (2005). Computers that care : Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, 62 :161–178.
- [DeRosis et al., 2003] DeRosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., and Carolis, B. D. (2003). From greta's mind to her face : Modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, 59(1-2) :p. 81–118.
- [Dyer, 1987] Dyer, M. G. (1987). Emotions and their computations : three computer models. *Cognition and Emotion*, 1(3) :323–347.
- [Helmut and Mitsuru, 2005] Helmut, P. and Mitsuru, I. (2005). The empathic companion : A character-based interface that addresses users' affective states. *International Journal of Applied Artificial Intelligence*, 19 :297–285.
- [Johnson et al., 2000] Johnson, W., Rickel, J., and Lester, J. (2000). Animated pedagogical agents : Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11 :47–78.
- [Lazarus, 2001] Lazarus, R. S. (2001). Relational meaning and discrete emotions. In Scherer, K., Schorr, A., and Johnstone, T., editors, *Appraisal Processes in Emotion : Theory, Methods, Research*, pages 37–69. Oxford University Press.
- [Meyer, 2006] Meyer, J. (2006). Reasoning about emotional agents : Research articles. *International Journal of Intelligent Systems*, 21(6) :601–619.
- [Ochs et al., 2006] Ochs, M., Pelachaud, C., and Sadek, D. (2006). Les conditions de déclenchement des émotions d'un agent empathique. In *Workshop Francophone sur les Agents Conversationnels Animés (WACA)* (<http://www.irit.fr/WACA/>).
- [Ortony et al., 1988] Ortony, A., Clore, G., and Collins, A. (1988). *The cognitive structure of emotions*. Cambridge University Press, United Kingdom.
- [Pacherie, 2004] Pacherie, E. (2004). L'empathie et ses degrés. In Berthoz, A. and Jorland, G., editors, *L'empathie*, pages 149–181. Editions Odile Jacob.
- [Picard, 1997] Picard, R. (1997). *Affective Computing*. MIT Press.
- [Poggi, 2004] Poggi, I. (2004). Emotions from mind to mind. In *Proceedings of the Workshop on Empathic Agents*. AAMAS, pages 11–17.
- [Roseman, 2001] Roseman, I. J. (2001). A model of appraisal in the emotion system. In Klaus Scherer, Angela Schorr, T. J., editor, *Appraisal Processes in Emotion : Theory, Methods, Research*, pages 68–91. Oxford University Press.
- [Sadek, 1991] Sadek, D. (1991). *Attitudes mentales et interaction rationnelle : vers une théorie formelle de la communication*. PhD thesis, Université Rennes 1.
- [Sadek, 1992] Sadek, D. (1992). A study in logic of intention. In *Proceeding of the 3rd International Conference on Principles of Knowledge Representation and Reasoning (KR'92)*.
- [Scherer, 2000] Scherer, K. (2000). Emotion. In Hewstone, M. and Stroebe, W., editors, *Introduction to Social Psychology : A European perspective*, pages 151–191. Oxford Blackwell Publishers, Oxford.