Un modèle logique de théorie de l'esprit pour un agent virtuel dans le contexte de simulation d'entretien d'embauche

M Belkaid, Nicolas Sabouret

Un modèle logique de théorie de l'esprit pour un agent virtuel dans le contexte de simulation d'entretien d'embauche

M. Belkaïd[1] N. Sabouret[2]

[1] ETIS, UMR 8051, Cergy
[2] LIMSI-CNRS, UPR 3251, Orsay

Résumé

Dans le contexte de la simulation d'entretiens d'embauche, le recruteur virtuel doit être capable de se représenter et de raisonner sur les états mentaux de l'utilisateur en s'appuyant sur des indices non-verbaux qui sont des indices de ses émotions et de son attitude sociale. Dans cet article, nous proposons un modèle formel de théorie de l'esprit (ToM) pour des agents virtuels dans le contexte d'interaction humain-agent, en nous concentrant sur les dimensions affectives des interactions. Ce modèle combine deux paradigmes de théorie de l'esprit et s'appuie sur une logique modale de type BDI dans laquelle sont décrites les règles d'inférence sur les états mentaux, les émotions et les relations sociales entre les acteurs. Nous présentons les résultats d'une étude préliminaire sur l'impact d'un tel modèle dans le contexte de la simulation d'entretien d'embauches.

Mots Clef

Théorie de l'esprit, modèles cognitifs, approches logiques, agents virtuels intelligents.

Abstract

In job interview simulation, the virtual interviewer must be capable of representing and reasoning about the user's mental state based on social cues that inform the system about his/her affects and social attitude. In this paper, we propose a formal model of Theory of Mind (ToM) for virtual agent in the context of human-agent interaction that focuses on the affective dimension. It relies on a hybrid ToM that combines the two major paradigms of the domain. Our framework is based on modal logic and inference rules about the mental states, emotions and social relations of both actors. Finally, we present preliminary results regarding the impact of such a model on natural interaction in the context of job interviews simulation.

Keywords

Theory of Mind, Cognitive Models, Logic-Based Approaches, Intelligent Virtual Agents.

1 Introduction

Le travail présenté dans cet article se situe dans le contexte de l'utilisation d'agents virtuels pour la simulation d'entretien d'embauche, qui a reçu un intérêt croissant de la communauté ces dernières années [2, 5, 16]. En effet, sur le plan sociétal, l'aide à l'insertion professionnel est un objectif majeur de nos sociétés (le taux de chômage chez les moins de 25 ans en Europe a dépassé 25%) et il a été montré que la simulation d'entretien d'embauche, en particulier avec des agents virtuels, peut à améliorer la confiance en soi et les compétences sociales des jeunes [8, 16, 26]. Sur le plan théorique, la simulation d'entretien d'embauche est une situation contrôlée dans laquelle il semble plus facile d'étudier la reconnaissance, le raisonnement et la synthèse de comportements affectifs, qui sont des problèmes trop difficiles pour être abordés dans un cadre général.

Notre objectif est de faire des agents dont la réaction verbale et non-verbale est cohérente avec les entrées non-verbales (sourires, expressions émotionnelles, mouvements du corps). Alors que la majorité des modèles comportementaux pour les agents virtuels proposent des modèles plutôt réactifs[16, 20, 21], nous proposons de raisonner sur les états mentaux de l'interlocuteur pour adapter le comportement des agents. La théorie de l'esprit (ou ToM, pour *Theory of Mind*) est la capacité qu'ont les humains et les primates à interpréter, prédire et même influence le comportement des autres[4]. Dans le contexte d'agents intelligents pour la pratique de compétences sociales, cette capacité nous semble être la clef vers des comportements plus réalistes.

Dans la section suivante, nous présentons brièvement les recherches qui sont à la base de nos travaux. Les sections 3 présentent l'architecture générale et notre modèle logique de ToM. La section 4 présente notre implémentation dans le contexte de la simulation d'entretiens d'embauche. Nous présentons les grandes lignes d'une expérimentation préliminaire dans la section 5 et nous discutons des résultats et des perspectives dans la section 5.2.

2 Travaux connexes

Les modèles d'appraisal comme CPM [25] ou OCC [19] peuvent être utilisés pour permettre aux agents virtuels de raisonner sur la dimension affective de l'interaction, portée par le comportement non-verbal des deux interlocuteurs. Ainsi, [1, 10] proposent des implémentations BDI de OCC. Le modèle de double appraisal proposé dans FAtiMA [3],

bien qu'il ne repose pas sur un modèle logique, est un premier exemple de théorie de l'esprit basée sur OCC. Notre objectif est de définir un modèle logique de théorie de l'esprit orienté vers les émotions, en s'appuyant sur les modèles logiques BDI comme [15] et [10] qui proposent une formalisation du raisonnement sur les états mentaux de l'interlocuteur.

En sciences humaines, un débat subsiste sur la nature des mécanismes de théorie de l'esprit : les défenseurs de la *theory-theory*(TT) postulent que la ToM s'appuie sur des règles de sens commun[7], alors que les partisans de la *simulation-theory* (ST) [12] defendent l'idée d'une projection dans l'état mental de l'interlocuteur. Plusieurs travaux ont montré qu'aucune de ces deux visions n'étaient suffisante [27], ce qui a donné naissance à des approches hybrides [7][12].

Les modèles informatique de la ToM choisissent en général l'une ou l'autre des approches. Ainsi, [3] repose sur une approche ST alors que [6, 23] se situent dans une approche TT. Lorsqu'elles sont combinées (comme dans [14]), elles sont implémentées de manière disjointes. Notre proposition est de définir un modèle logique qui combine de manière naturelle ces deux approches au sein d'un même moteur de raisonnement.

3 Architecture et modèle logique

Notre architecture, illustrée sur la figure 1, est composée des éléments suivants. Les états mentaux sont les croyances, attitudes, buts et intentions des agents. Les croyances portent sur des faits, des règles du monde (au sens de la TT) et sur les états mentaux des autres agents. Les attitudes décrivent l'évaluation de l'état du monde et, par extension, ses buts. Dans notre modèle les intensions ne portent que sur la prochaine action. Le moteur d'inférence comprend un modèle délibératif de type *folk-psychology* (au sens de la ST) qui permet de mettre à jour les croyances de l'agent, un modèle de raisonnement de sens commun (au sens de la TT, dans laquelle on retrouve des règles spécifiques au domaine, la simulation d'entretien d'embauche dans notre cas) et enfin le modèle affectif basé sur OCC.

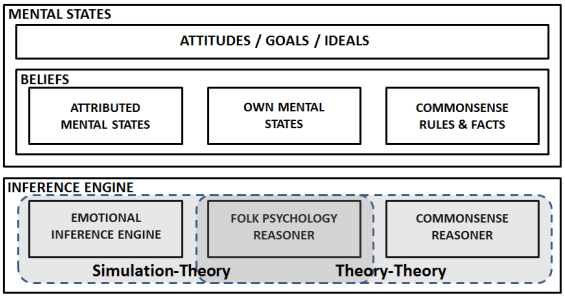

FIGURE 1 – Notre architecture hybride de ToM.

3.1 Syntaxe du modèle logique

Soit ATM un ensemble de propositions décrivant des faits (par exemple : "le salaire proposé est élevé"), ACT un ensemble d'action (par exemple : "se présenter"), ILL un ensemble d'actes de langages, AGT un ensemble d'agents (l'agent virtuel, son interlocuteur et éventuellement d'autres individus), EMO un ensemble de catégories d'émotions. Nous appelons *événements* les actions dans lesquelles l'un des interlocuteurs prend part, comme dans [18]. Un événement $e \in EVT$ est un tuple dans $AGT \times AGT \times (ACT \cup ILL(ATM))$ représentant l'agent qui effectue l'action, celui qui la subit et enfin l'action elle-même, qui peut être aussi un acte de langage (par exemple : "dire que *le salaire est élevé*"). Nous y ajoutons un degré de plausibilté comme cela se fait habituellement en BDI.
Notre langage est défini par la grammaire suivante :

$$\begin{aligned} &Evt : \epsilon ::= \langle a, (a|\varnothing), \alpha\rangle \mid \langle a, a, Spk(\varsigma, \varphi)\rangle \\ &Prp : \pi ::= p \mid \epsilon \mid Like^k_{a,b} \mid Dom^k_{a,b} \\ &Fml : \varphi ::= \pi \mid Bel^l_a(\varphi) \mid Att^k_a(\varphi) \mid Int_a(\varphi) \mid \\ &\qquad Emo^i_{a,(b|\varnothing)}(\varepsilon, \varphi) \mid N(\varphi) \mid U(\varphi, \varphi) \mid \neg\varphi \mid \varphi \wedge \varphi \end{aligned} \tag{1}$$

avec $a, b \in AGT$, $\alpha \in ACT$, $p \in ATM$, $\epsilon \in EVT$, $\varepsilon \in EMO$, $\varsigma \in ILL$, $l, i \in [0, 1]$, $k \in [-1, 1]$. $Like$, Dom, Bel, Att et Int sont des opérateurs de la logique modale et N et U sont les opérateurs temporels *Next* et Until de la logique temporelle LTL et CTL* [22]. Les autre opérateur temporels F et G ainsi que les opérateurs booléens $\top$, $\bot$, $\vee$ et $\Rightarrow$ sont définis de manière classique. De plus, dans la description des événements, nous autorisons l'utilisation de "$-$" pour désigner n'importe quel atome.
Notre modèle de relation sociale est basé sur [17] : $Like^k_{a,b}$ détermine le degré d'appréciation et $Dom^k_{a,b}$ le degré de dominance.
$Bel^l_a(\varphi)$ décrit une croyance, comme dans [10], et se lit "l'agent a croit que φ avec une certitude l". De même, $Att^k_a(\varphi)$ décrit une attitude. Dans notre modèle, cet opérateur sera utilisé pour décrire des désirs, des idéaux et des buts, qui seront représentés avec leurs propres opérateurs modaux comme dans [1, 13].
Dans la suite, nous utiliserons l'opérateur $\stackrel{\text{def}}{=}$ pour la définition de nouveaux opérateurs, alors que l'opérateur $\stackrel{\text{def}}{\Longrightarrow}$ sera utilisé pour décrire nos règles d'inférences. Ainsi :

$$\begin{aligned} Des^k_a(\varphi) &\stackrel{\text{def}}{=} Att^k_a(F(\varphi)) \\ Ideal^{k>0}_a(\varphi) &\stackrel{\text{def}}{=} Att^{k>0}_a(G(\varphi)) = Des^{-k<0}_a(\neg\varphi) \end{aligned} \tag{2}$$

Un désir est quelque chose envers lequel l'agent a une attitude positive, alors qu'un idéal est quelque chose que l'agent souhaiterait toujours vrai. L'objet d'une attitude, d'un désir ou d'un idéal peut être un atome ("préserver la forêt") ou une formule plus complexe comme une croyance ou un opérateur temporel.

Comme en BDI [24], nous notons $Int_a(\varphi)$ les intentions (plans) d'un agent.
$Emo^i_{a,(b|\varnothing)}(\varepsilon, \varphi)$ représente les émotions. Conformément à [11], une émotion ε est toujours à propos d'un fait $varphi$ et peut être dirigée vers un agent b, avec $\varepsilon \in EMO$ la catégorie émotionnelle et i l'intensité.
Nous introduisons l'opérateur $Resp_a$ pour décrire la responsabilité directe (contrairement à [1, 13], nous ne considérons pas le cas où l'agent est responsable d'une situation qu'il aurait pu éviter) :

$$Resp_a(\epsilon) \stackrel{\text{def}}{=} (\epsilon = \langle a, -, - \rangle) \tag{3}$$

3.2 Semantique

Notre sémantique est basé sur la théorie des mondes possibles. Soit $\mathcal{F} = \langle \mathcal{W}, \mathcal{B}, \mathcal{D}, \mathcal{I}, \mathcal{E} \rangle$ avec :

- W l'ensemble non-vide des mondes possibles,
- $\mathcal{B} : AGT \rightarrow (W \rightarrow 2^W)$ la fonction qui associe à chaque agent $a \in AGT$ et à chaque monde $w \in W$ l'ensemble des mondes accessibles par les croyances $\mathcal{B}_a(w)$,
- $\mathcal{D} : AGT \rightarrow (W \times [0,1] \rightarrow 2^W)$ la fonction qui associe à chaque agent $a \in AGT$ et à chaque monde $w \in W$ avec un degré de désirabilité l l'ensemble des mondes accssibles par les désirs $\mathcal{D}_a(w, l)$,
- $\mathcal{I} : AGT \rightarrow (W \rightarrow 2^W)$ la function qui associe à chaque agent $a \in AGT$ et à chaque monde $w \in W$ l'ensembles des mondes accessibles par l'intention $\mathcal{I}_a(w)$,
- $\mathcal{E} : EVT \rightarrow W$ la fonction qui associe à chaque évenement $\epsilon \in EVT$ le monde résultant.

Soit un modèle $\mathcal{M} = \langle \mathcal{F}, \mathcal{V} \rangle$ (avec $\mathcal{V} : W \rightarrow ATM$ une fonction d'évaluation). Nous notons $\mathcal{M}, w \models \varphi$ le fait que φ est vrai dans w. Les valeurs de vérités sont définies de manière classique par induction :

- $\mathcal{M}, w \models p$ ssi $p \in \mathcal{V}(w)$;
- $\mathcal{M}, w \models \neg\varphi$ ssi $\mathcal{M}, w \models \varphi$ n'est pas vrai ;
- $\mathcal{M}, w \models \varphi \wedge \psi$ ssi $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$;
- $\mathcal{M}, w \models Bel^l_a(\varphi)$ ssi $\frac{card(\mathcal{GB}_a(w))}{card(\mathcal{B}_a(w))} = l$
  avec $\mathcal{GB}_a(w) = \{v \in \mathcal{B}_a(w) ; \mathcal{M}, v \models \varphi\}$;
- $\mathcal{M}, w \models Des^l_a(\varphi)$ ssi $\mathcal{M}, v \models \varphi \; \forall v \in \mathcal{D}_a(w, l)$;
- $\mathcal{M}, w \models Int_a(\varphi)$ ssi $\mathcal{M}, v \models \varphi \; \forall v \in \mathcal{I}_a(w)$;
- $\mathcal{M}, w \models \epsilon$ ssi $\mathcal{M}, v \models \top \; \forall v \in \mathcal{E}(\epsilon)$;

Dans la prochaine section, nous présentons quelques règles d'inférence de notre modèle. Pour des raisons de place, toutes les règles ne peuvent pas être décrites dans cet article mais nous présentons les plus significatives. Lorsque cela sera nécessaire, nous noterons f la combinaison des degrés de croyances, d'intensité émotionnelle, etc qui sera décrite dans la section 4.

3.3 Quelques règles

Comme dans [10, 1], les relations de croyance $\mathcal{B}$ sont transitive et euclidiennes :

$$Bel^l_a(\varphi) \stackrel{\text{def}}{\Longrightarrow} Bel^1_a(Bel^l_a(\varphi)) \tag{4}$$

Ainsi, les agents sont conscients de leurs propres états mentaux et de leurs relations sociales. Cependant, choisissons que $\mathcal{B}$ ne forme pas une relation binaire :

$$Bel^l_a(\varphi) \stackrel{\text{def}}{\Longrightarrow} Bel^{1-l}_a(\neg\varphi) \tag{5}$$

Nous définissons $0.5 < mod\_th < str\_th$ deux limites pour représenter le fait que l'agent croit peu ($l < mod\_th$), moyennement ($mod\_th < l < str\_th$) ou fortement ($l > str\_th$) quelque chose. Enfin, nos agents sont capables de déductions :

$$Bel^l_a(\psi) \wedge Bel^{l'}_a(\psi \Rightarrow \varphi) \stackrel{\text{def}}{\Longrightarrow} Bel^{f(l,l')}_a(\varphi) \tag{6}$$

Les attitudes peuvent être positives ou négatives mais on les suppose consistants :

$$\mathcal{M}, w \models (Att^k_a(\varphi) \wedge Att^{k'}_a(\neg\varphi)) \text{ iff } k = -k' \tag{7}$$

Toutefois, un agent peut désirer quelque chose qui conduit à quelque chose de non-désiré. C'est au moment du choix d'action que ce conflit sera géré :

$$Des^k_a(\varphi) \wedge Bel^{l>str\_th}_a(\psi \Rightarrow F(\varphi)) \wedge \neg IncDes^k_a(\psi) \stackrel{\text{def}}{\Longrightarrow} N(Des^k_a(\psi)) \tag{8}$$

avec $IncDes^k_a(\varphi)$ le désir inconsistant :

$$IncDes^k_a(\varphi) \stackrel{\text{def}}{=} (Bel^{l>str\_th}_a(\varphi \Rightarrow \neg\psi) \wedge Des^{k'>0}_a(\psi)) \vee (Bel^{l>str\_th}_a(\varphi \Rightarrow \psi) \wedge Des^{k'<0}_a(\psi)) \tag{9}$$

Ainsi, le désir φ est inconsistant si l'agent croit fortement qu'il conduit à un fait indésirable ψ. Conformément au modèle BDI [24], nous définissons les buts comme des désirs consistants et qu'il croit atteignables. Pour cela, nous introduisons la limite des_th :

$$Goal^{k>0}_a(\varphi) \stackrel{\text{def}}{=} Des^{k>des\_th}_a(\varphi) \wedge Bel^l_a(F(\varphi)) \wedge \neg IncDes^k_a(\varphi) \tag{10}$$

Enfin, un but est transformé en intention lors que l'agent peut l'atteindre :

$$Goal^{k>0}_a(\epsilon) \wedge Resp_a(\epsilon) \stackrel{\text{def}}{\Longrightarrow} N(Int_a(\epsilon)) \tag{11}$$

ou, comme dans [6], parce qu'il croit qu'il existe un moyen de l'atteindre :

$$Goal^{k>0}_a(\varphi) \wedge Bel^{l>str\_th}_a(\psi \Rightarrow F(\varphi)) \wedge \neg IncDes^k_a(\psi) \wedge Bel^{l'}_a(F(\psi)) \stackrel{\text{def}}{\Longrightarrow} N(Int_a(\psi)) \tag{12}$$

Lors qu'un agent a une intention et peut la réaliser, il le fait :

$$Int_a(\varphi) \wedge Bel^{l>str\_th}_a(\psi \Rightarrow F(\varphi)) \stackrel{\text{def}}{\Longrightarrow} Int_a(\psi) \qquad Int_a(\epsilon) \wedge Resp_a(\epsilon) \stackrel{\text{def}}{\Longrightarrow} N(\epsilon) \tag{13}$$

Dans notre modèle, comme dans [18, 9, 3], les attitudes sont influencées non seulement par les croyances, mais aussi par la relation sociale :

$$\begin{aligned} Bel_a^{l>str\_th}(\varphi) \wedge Att_a^k(F(\varphi)) \wedge Bel_a^{l'}(Att_b^{k'}(F(\varphi))) \\ \wedge Like_{a,b}^h \wedge Dom_{a,b}^{h'} \stackrel{\text{def}}{\Longrightarrow} Att_a^{f(k,k',h,h')}(\varphi) \\ Bel_a^{l>str\_th}(Des_b^k(\varphi)) \wedge Like_{a,b}^{k'>0} \stackrel{\text{def}}{\Longrightarrow} N(Des_a^{f(k,k')}(\varphi))) \end{aligned} \tag{14}$$

Enfin, nos règles d'appraison sont conformes à ce qui se fait classiquement dans la litérature [1, 13, 10]. Par exemple :

$$\begin{aligned} Bel_a^l(\gamma) \wedge Att_a^{k>0}(\gamma) \stackrel{\text{def}}{\Longrightarrow} N(Joy_a^{i=f(l,k)}(\gamma)) \\ Bel_a^l(F(\gamma)) \wedge Des_a^{k<0}(\gamma) \stackrel{\text{def}}{\Longrightarrow} N(Fear_a^{i=f(l,k)}(\gamma)) \\ Bel_a^l(\gamma) \wedge Ideal_a^k(\gamma) \wedge Bel_a^{l'}(Rsp_b(\gamma)) \\ \stackrel{\text{def}}{\Longrightarrow} N(Admiration_{a,b}^{i=f(l,l',k)}(\gamma)) \end{aligned} \tag{15}$$

Enfin, les règles de sens commun dépendent du domaine. Dans le contexte de l'entretien d'embauche, nous aurons par exemple :

$$\begin{aligned} &Des_r^{0.77}(\neg Emo_{\_,c}^i(distress)) \wedge i > 0.5 \\ &Bel_r^{0.8}(Att_c^{-0.5} salary\_is\_bad) \end{aligned}$$

Pour représenter un agent qui ne souhaite pas rendre le candidat triste et qui pense que les candidats souhaitent généralement obtenir un bon salaire.

4 Implémentation

Le modèle théorique que nous avons présenté dans la section précédent est censé être indépendant du domaine, mis à part les règles de sens commun. Dans notre implémentation pour la simulation d'entretien d'embauche, dans le cadre du projet TARDIS [2]. L'intérêt de l'entretien d'embauche pour la validation de notre modèle est qu'il s'agit de situations de dialogue diadique semi-structurés où le recruteur a souvent la possibilité de raisonner sur les états mentaux et les émotions du candidat.

Notre cadre logique et le moteur d'inférence ont été impémentés en SWI-Prolog. Ce moteur a été couplé avec un programme C++ qui gère le tour de parole et la communication entre les modules. Conformément au modèle BDI, à chaque tour, l'agent interprète les émotions exprimées par son interlocuteur pour générer une liste d'actions possibles, en sélectionner une pour mettre à jour ses intentions et les exécuter.

La difficulté lors de l'implémentation de notre modèle est de fixer les limites pour les croyances et les buts, et de définir les fonctions de combinaison que nous avons noté f dans les formules de la section précédente. Dans notre implémentation, nous avons choisi de manière arbitraire de fixer : $mod\_th = 0.5$, $str\_th = 0.75$ et $des\_th = 0.7$.

Les fonctions de combinaison peuvent être regroupées en deux catégories :

– Pour la dynamique des attitudes et des croyances (par exemple l'équation 14), nous utilisons une simple moyenne suivie d'une normalisation :

$$f(k, k') = ((k + k')/4) + 0.5$$

– Pour les émotions (équation 15), nous combinons une influence linéaire de l'attitude (par exemple, la joie est linéairement corrélée à l'attitude envers l'objet de l'émotion) avec une influence logarithmique de degré de croyance. Ainsi, nous obtenons des émotions plus fortes avec des croyances relativement faibles :

$$f(l, k) = \frac{k}{2} \times \frac{Log(2l - 1) - min}{min} + 0.5$$

où min est la limite de $Log(x)$ lorsque $x \to 0$, soit la plus petite valeur possible dans l'ordinateur. Le facteur $2l - 1$ permet d'ajuster la valeur dans $[0, 1]$ avant le calcul de l'intensité, puis nous le réajustons dans $[0.5, 1]$ pour obtenir des intensités plus significatives.

Les règles de sens commun correspondant à la situation d'entretien d'embauche définissent l'ensemble des actes de dialogue (parler du salaire, de l'expérience) et des attentes en termes d'impact (dire au candidat qu'il est en retard devrait le mettre mal à l'aise). L'agent choisit alors les actions à effectuer, c'est-à-dire les actes de langage, en fonction de ses buts courants (en terme d'état affectif de l'interlocuteur et de sujets à aborder : par exemple, je souhaite mettre l'interlocuteur à l'aise mais je dois aborder la question du salaire). De plus, nous avons définis des opérateurs spécifiques pour décrire la confiance en soi, la motivation et la compétence professionnelle du candidat. Les degrés de croyances pour ces faits sont calculés en fonction de ses réactions émotionnelles aux questions de l'agent à l'aide de règles simple (de type TT). Par exemple, des hésitations lors de la réponse à une question sont un indice de non-confiance en soi.

La perception des états affectifs à travers le comportement non-verbal de l'interlocuteur est gérée par un autre module dans le projet TARDIS(voir [2]). Dans cet article, comme nous le verrons, nous supposons que nous avons des entrées numériques dont on ne sait pas comment elles ont été obtenues à partir de capteurs.

5 Évaluation préliminaire

Notre premier objectif était d'étudier l'impact d'un tel modèle de théorie de l'esprit sur la qualité ou la difficulté d'un entretien d'embauche avec un agent virtuel. Pour cela, nous avons évalué notre modèle en situation avec 30 sujets – 11 femmes et 19 hommes – issus du personnel de notre laboratoire, qui ont interagit à travers une interface graphique simplifiée avec notre modèle logique. Aucun agent virtuel

n'était présent : l'état mental du recruteur et ses croyances en terme de confiance en soi, de motivation et de compétence du candidat étaient représentés par des barres de progression (dont les valeurs allaient de -1 à 1). De même, aucun capteur n'était utilisé : les utilisateurs devaient saisir leur comportement à l'aide de 8 indicateurs (ascenseur à placer entre 0 et 1) : soulagé, embarrassé, hésitant, stressé, mal à l'aise, concentré, agressif, je m'ennuie. Le choix de ces indicateurs provient des recherches du projet TARDIS [2].

Les sujets devaient jouer le rôle du candidat à un poste de secrétaire et pouvaient adopter la personnalité qu'ils souhaitaient. à l'issue de l'entretien, ils devaient remplir un questionnaire de 11 items (sur une échelle de Lickert à 5 valeurs) portant sur la crédibilité des réactions et la qualité des évaluations faites par l'agent. Chaque sujet interagissait avec l'une des trois versions possible de l'agent : celui dont la base de connaissance avait pour objectif de mettre le candidat à l'aise (PROFIL_A), celui qui posait des questions classiques sans but précis sur l'état mental de l'utilisateur (PROFIL_B) et celui qui posait des questione embarrassantes (PROFIL_C). Les 3 agents utilisent le même moteur de raisonnement.

Notre première hypothèse est que la variation du profil aura un impact sur la réaction émotionnelle des candidats (du moins, telle qu'elle est exprimée par les 8 indicateurs manipulés par le sujet). Notre mesure porte sur la somme des intensités émotionnelles exprimées.

5.1 Resultats

Nos résultats montrent, suivant un test de Kruskal-Wallis, un effet principal du profil (PROFIL_X) sur la somme des intensités émotionnelles ($Chi^2(2, 629) = 11.435; p < 0.01$) et, en particulier, sur l'embarras exprimé ($Chi^2(2, 629) = 6.231; p < 0.05$) et sur la concentration exprimée ($Chi^2(2, 629) = 9.218; p < 0.01$). Cela signifie que le profil du recruteur a un impact sur les affects décrits (et peut-être exprimés en situation réelle) par les sujets. Un test de Mann-Whitney montre aussi que test les participants qui ont interagit avec le profil A (compréhensif) choisissent plus d'embarras et plus de concentraction, alors que ceux qui ont interagit avec le profil C (difficile) choisissent plus de stress, de "mal à l'aise" et de concentration. Nos résultats sont illustrés sur la figure 2.

5.2 Discussion

La théorie de l'esprit est un phénomène complexe qui fait intervenir d'autres processus cognitifs (dont la mémorisation, les capacités de raisonnement, etc) et perceptifs (*e.g.* l'interprétation des signaux sociaux). C'est pourquoi il est difficile non seulement de la modéliser, mais aussi de l'évaluer. Un protocole simple comme celui que nous avons présenté ici n'est pas suffisant pour évaluer l'impact de la ToM sur la qualité de l'entraînement à l'entretien d'embauche. Pour commencer, l'utilisation d'une interface graphique et non d'un agent virtuel ne permet pas aux utilisateurs de s'immerger dans la situation. Notre hypothèse est

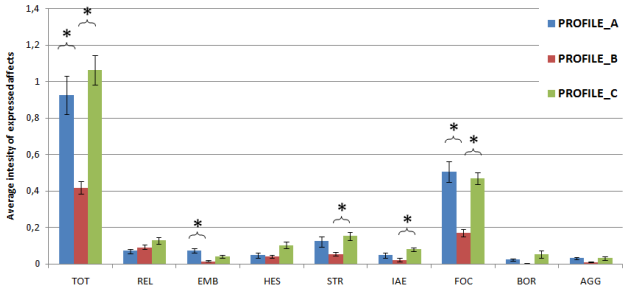

FIGURE 2 – Intensité moyenne et écart-type choisi par les participants en fonction du profil et des affects considérés. Les effets significatifs (*) sont sur les affects embarras, stress, mal à l'aise, concentré et sur la somme totale des intensités.

que le dispositif complet, tel qu'il est proposé dans le projet TARDIS, devrait permettre d'évaluer la réactivité et les processus de raisonnement de l'agent virtuel. Cependant, dans la litérature, il n'existe pas de protocole expérimental pour l'évaluation de la qualité d'une théorie de l'esprit dans une situation d'interaction. Les travaux les plus proches [14, 23] évaluent la capacité de la ToM artificielle à expliquer les choix dans le scénario en comparant les décisions avec le modèle de la tâche. Mais ces modèles ne portent que sur la partie verbale de l'interaction, alors que notre ToM s'intéresse à la composante co-verbale (par la sélection d'actes expressifs).

Cependant, notre étude préliminaire nous a permis de mettre en évidence quelques résultats pour l'évaluation d'un tel modèle dans un contexte de simulation d'entretien. Tout d'abord, le fait que le profil B (neutre) n'utilise pas de théorie de l'esprit pour sélectionner les questions qu'il pose conduit à des réactions affectives moins importantes. Il semblerait donc que la ToM sur les émotions dans le contexte de l'interaction dialogique ait un réel impact sur la réaction de l'interlocuteur. De plus, cette étude montre la nécessité de disposer de différents profils de recruteurs (au niveau du modèle de raisonnement, pas au niveau de l'apparence ou du comportement non-verbal) pour éliciter des émotions différentes.

L'utilisation d'un modèle logique présente un autre intérêt majeur pour l'entraînement des candidats à l'emploi : chaque déduction peut être retrouvée, expliquée et fournie à l'utilisateur, pour lui permettre de mieux comprendre l'impact de ses réactions sur les états mentaux de l'agent virtuel, et l'interprétation qu'il a fait de l'état mental du candidat. C'est pourquoi nous poursuivons actuellement nos travaux pour l'intégration de modèles logiques de ToM au sein d'agents virtuels.

Références

[1] C. Adam, A. Herzig, and D. Longin. A logical formalization of the OCC theory of emotions. *Synthese*, 168(2) :201–248, Feb. 2009.

[2] K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, H. Jones, M. Ochs, C. Pelachaud, K. Porayska-Pomsta, R. Paola, and N. Sabouret. The TARDIS framework : intelligent virtual agents for social coaching in job interviews. *Proceedings of the Tenth International Conference on Advances in Computer Entertainment Technology (ACE-13). Enschede, the Netherlands. LNCS 8253*, page in press, 2013.

[3] R. Aylett and S. Louchart. If I were you : double appraisal in affective agents. In *Proceedings of the 7th international joint conference on Autonomous Agents and MultiAgent Systems*, pages 1233–1236, 2008.

[4] S. Baron-Cohen. *Mindblindness : An essay on autism and theory of mind*. MIT press, 1997.

[5] L. M. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer. Cicero - towards a multimodal virtual audience platform for public speaking training. In *Proc. 2013 International Conference on Intelligent Virtual Agents*, pages 116–128, 2013.

[6] T. Bosse, Z. A. Memon, and J. Treur. A recursive BDI agent model for Theory of Mind and its applications. *Applied Artificial Intelligence*, 25(1) :1–44, 2011.

[7] G. Botterill and P. Carruthers. *The philosophy of psychology*. Cambridge University Press, 1999.

[8] J. Bynner and S. Parsons. Social Exclusion and the Transition from School to Work : The Case of Young People Not in Education, Employment, or Training (NEET). *Journal of Vocational Behavior*, 60(2) :289–309, Apr. 2002.

[9] C. Castelfranchi. Modelling social action for AI agents. *IJCAI'97 Proceedings of the Fifteenth international joint conference on Artifical intelligence - Volume 2*, 103(1) :1567–1576, 1997.

[10] M. Dastani and E. Lorini. A logic of emotions : from appraisal to coping. In *Proceedings of the 11thiInternational conference on Autonomous Agents and Multiagent Systems*, pages 1133–1140, 2012.

[11] N. H. Frijda. *The emotions*. Cambridge University Press, 1986.

[12] A. I. Goldman. *Simulating minds : The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press, 2006.

[13] N. Guiraud, D. Longin, E. Lorini, S. Pesty, and J. Rivière. The face of emotions : a logical formalization of expressive speech acts. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, pages 1031–1038, 2011.

[14] M. Harbers. Explaining agent behavior in virtual training. *SIKS dissertation series*, 2011(35), 2011.

[15] A. Herzig and D. Longin. A logic of intention with cooperation principles and with assertive speech acts as communication primitives. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems : part 2*, pages 920–927. ACM, 2002.

[16] M. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. Picard. MACH : My Automated Conversation coacH. In *Proc. 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM Press, 2013.

[17] T. F. Leary et al. *Interpersonal diagnosis of personality*. Ronald Press New York, 1957.

[18] M. Ochs, N. Sabouret, and V. Corruble. Simulation of the Dynamics of Nonplayer Characters' Emotions and Social Relations in Games. *Computational Intelligence and AI in Games, IEEE Transactions on*, 1(4) :281–297, 2009.

[19] A. Ortony, G. L. Clore, and A. Collins. The Cognitive Structure of Emotions, 1990.

[20] A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobreperez, S. Woods, C. Zoll, and L. Hall. Caring for agents and agents that care : Building empathic relations with synthetic agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 194–201, Washington, DC, USA, 2004. IEEE Computer Society.

[21] L. Pareto, D. Schwartz, and L. Svensson. Learning by guiding a teachable agent to play an educational game. *in Education Building Learning*, pages 1–3, 2009.

[22] A. Pnueli. The temporal logic of programs. In *Foundations of Computer Science, 1977., 18th Annual Symposium on*, pages 46–57. IEEE, 1977.

[23] D. V. Pynadath, N. Wang, and S. C. Marsella. Are you thinking what I'm thinking ? An Evaluation of a Simplified Theory of Mind. In *Intelligent Virtual Agents*, pages 44–57. Springer, 2013.

[24] A. S. Rao and M. P. Georgeff. Modeling Rational Agents within a BDI-Architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, 1991.

[25] K. R. Scherer. Emotion and emotional competence : conceptual and theoretical issues for modelling agents. *Blueprint for Affective Computing*, pages 3–20, 2010.

[26] M. Sieverding. 'Be Cool !' : Emotional costs of hiding feelings in a job interview. *International Journal of Selection and Assessment*, 17(4), 2009.

[27] K. Vogeley, P. Bussfeld, A. Newen, S. Herrmann, F. Happe, P. Falkai, W. Maier, N. J. Shah, G. R. Fink, and K. Zilles. Mind reading : neural mechanisms of theory of mind and self-perspective. *Neuroimage*, 14(1) :170–181, 2001.