First seconds matter : Mangaing first impressions for a more engaging virtual agent

Béatrice Biancardi

EDITE - ED 130

THÈSE DE DOCTORAT DE
SORBONNE UNIVERSITÉ

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)
présentée et soutenue publiquement par

Beatrice Biancardi

le 8 juillet 2019

Les premières secondes comptent :

Gérer les premières impressions pour un agent virtuel plus engageant.

Directrice de thèse: **Catherine Pelachaud**

devant le jury composé de :

M. Pierre DE LOOR, Professeur, LabSTICC, ENIB	Rapporteur
M. Charles TIJUS, Professeur, ChArt, Université Paris 8	Rapporteur
M. Jean Julien AUCOUTURIER, CR, IRCAM-STMS, Sorbonne Université	Examinateur
Mme Laurence CHABY, Maître de conférences, Université Paris Descartes	Examinateur
M. Dirk HEYLEN, Professeur, University of Twente	Examinateur
M. Jean-Claude MARTIN, Professeur, LIMSI, Université Paris Sud	Examinateur
Mme Catherine PELACHAUD, DR, CNRS-ISIR, Sorbonne Université	Directrice de Thèse

Résumé Long

L’ interaction est un besoin fondamental de l’être humain. Dans notre vie quotidienne il y a de nombreuses occasions d’interagir avec différentes personnes, qu’il s’agisse d’inconnus ou de personnes très intimes, comme un partenaire ou un membre de la famille. Lorsque nous rencontrons des inconnus, les premiers moments sont critiques, car nous nous forgeons souvent des impressions sur l’autre, qui peuvent avoir des conséquences importantes comme un succès à un entretien d’embauche ou le fait de rencontrer de nouveau un partenaire potentiel (Ambady and Skowronski, 2008).

Goffman et al. (1978) définissent la *formation d’impression* comme le processus de perception, organisation et intégration de l’information afin de se forger des impressions cohérentes des autres (par exemple, en termes de personnalité et d’attitudes interpersonnelles). Nous, en tant qu’êtres humains, sommes conscients de ces mécanismes et essayons souvent de contrôler l’impression que les autres se forgent sur nous. Ce dernier processus s’appelle *gestion de l’impression* (Goffman et al., 1978), et concerne principalement le contrôle de l’apparence visuelle (par exemple, le type de coiffure et de vêtements). Cependant, nous essayons aussi de contrôler notre comportement social, mais il peut être difficile d’avoir un contrôle total sur tous les comportements qui se manifestent pendant l’interaction. En particulier, les comportements non verbaux sont cruciaux parce qu’ils peuvent révéler avec une grande précision une variété d’informations nous concernant, notamment l’orientation sexuelle (Ambady and Skowronski, 2008), la personnalité et les attitudes interpersonnelles (Rosenberg et al., 1968).

Au cours des dernières décennies, les interfaces anthropomorphes, comme les robots humanoïdes et les personnages virtuels, ont été de plus en plus utilisées dans plusieurs rôles, tels que des assistants pédagogiques, des compagnons, des formateurs. Lorsque l’on conçoit des Agents Conversationnels Animés (ACA), qui sont des personnages virtuels anthropomorphes capables d’interagir avec les utilisateurs en utilisant des comportements verbaux et non verbaux comme les gestes, les expressions faciales et la parole (pour plus de détails, voir Cassell (2000)), il est très important de prendre en compte leur perception pendant l’interaction. Les agents virtuels devraient être dotés de la capacité de maintenir des interactions engageantes avec les utilisateurs (Sidner and Dzikovska, 2005). Cela faciliterait la transmission de l’information par un guide virtuel, garantirait un changement de comportement pour un coach virtuel et créerait une relation avec un compagnon virtuel.

Comme dans les interactions humain-humain, les premiers moments d’une interaction avec les ACA sont critiques, car les utilisateurs forment des impressions sur eaux, ce qui peut affecter le reste de l’interaction, en termes d’engagement et de volonté de poursuivre celle-là(Cafaro et al., 2016). En gérant les comportements non verbaux d’un agent virtuel, nous pouvons améliorer la première impression que celui donne à l’utilisateur. Dans cette thèse, nous utilisons le terme “impression de l’agent” pour désigner l’impression donnée par l’agent afin d’être perçue par l’utilisateur avec différents niveaux de chaleur et de compétence (nous introduisons ces variables au paragraphe suivant). Nous utilisons le

terme “impression de l’utilisateur” pour désigner la représentation mentale de la chaleur et de la compétence de l’agent par cet utilisateur.

Le but de cette thèse est de construire une ACA capable de donner la meilleure première impression possible à l’utilisateur, en l’engageant ainsi efficacement dans une interaction. Cet objectif a été atteint en construisant une boucle interactive qui lie le comportement de l’agent à la réaction de l’utilisateur face à lui en temps réel. Nous nous sommes concentrés sur l’identification et la modélisation du comportement non verbal, à travers l’identification, la gestion et le maintien d’impressions de deux dimensions sociocognitives importantes dans les premières minutes d’une interaction avec un utilisateur, c’est-à-dire la chaleur et la compétence (C&C, (Fiske et al., 2007)).

Les travaux présentés dans cette thèse ont été réalisés dans le cadre du projet ANR IMPRESSIONS, en collaboration avec le Groupe Multimodal Interaction de l’Université de Genève. Elle a inclus aussi une collaboration externe avec le professeur Maurizio Mancini et une collaboration interne avec Paul Lerner et Soumia Dermouche.

Questions de Recherche

La motivation du travail présenté dans cette thèse était d’améliorer la qualité des impressions de l’agent générées dans l’utilisateur et l’engagement de l’utilisateur dans l’interaction humain-agent. En particulier, nous nous sommes concentrés sur l’agent virtuel et avons abordé les questions de recherche suivantes :

- (**RQ1**) *Comment **modéliser** les comportements non verbaux liés à la chaleur et à la compétence dans un agent virtuel?*
- (**RQ2**) *Comment **adapter** les comportements de l’agent virtuel aux impressions formées par les utilisateurs?*

Le projet ANR IMPRESSIONS a également abordé la question de recherche suivante: “*Comment **mesurer** les impressions à partir des expressions comportementales des utilisateurs?*”. Cela a été le sujet de thèse de Chen Wang. Nous avons collaboré avec elle et intégré son modèle pour détecter les impressions de l’utilisateur sur l’agent durant une interaction en temps réel. Ce travail est décrit dans le chapitre 9.

Pour aborder **RQ1** et **RQ2**, notre approche a commencé par examiner si les mêmes processus qui caractérisent la cognition sociale humaine s’appliquent également à l’interaction humain-agent. Nous sommes partis de l’analyse d’un corpus d’interactions naturelles humain-humain, visant à trouver des indices non verbaux provoquant différents degrés de C&C, dans le but d’appliquer ces résultats à l’interaction humain-agent. Notre approche était centrée sur le comportement non verbal, puisque cette modalité est très importante dans la formation d’impression et il est possible de la contrôler et de la manipuler dans un ACA.

Notre approche a suivi 4 étapes principales :

Étape 1 : Impressions de C&C dans l’interaction humain-humain

La première étape a consisté à analyser un corpus d’interactions dyadiques de partage de connaissances entre experts et novices, dans le but de construire un répertoire de signaux non verbaux suscitant différents degrés d’impressions de C&C. Cette étape visait à répondre à ces questions spécifiques :

- (**q1$_a$**) *Le comportement non verbal peut-il affecter les impressions de C&C?*
- (**q1$_b$**) *Si oui, quels sont les indices non verbaux associés à ces impressions?*

Ce travail est décrit en détail dans le chapitre 5.

Étape 2 : Impressions de C&C dans la perception d'un agent virtuel

Dans un deuxième temps, les résultats de la précédente étude ont été implémentés dans un ACA et manipulés afin de déterminer s'ils étaient perçus de la même manière lorsque réalisés par un ACA. Les jugements des utilisateurs sur la C&C ont été recueillis par le biais de questionnaires. Cette étape visait à répondre à ces questions spécifiques:

- (**q2$_a$**) *Un agent virtuel est-il perçu différemment en termes de C&C en fonction des comportements non-verbaux qu'il réalise?*
- (**q2$_b$**) *Si oui, quels sont les indices non verbaux (ou les combinaisons d'indices non verbaux) qui permettent de mieux percevoir l'agent en termes de C&C ?*
- (**q2$_c$**) *Est-ce que nos attentes et a priori d'un ACA influencent les impressions que nous nous forgeons ensuite sur lui?*

Ce travail est décrit en détail dans le chapitre 6.

Étape 3 : Architecture d'un système pour la gestion d'impressions de l'agent

En se basant sur les résultats des étapes précédentes, le but de cette étape était de répondre à **RQ2** en développant un système pour gérer les impressions de l'agent en temps réel lors de l'interaction avec un utilisateur. L'architecture se compose de 3 modules principaux: (1) un pour détecter les réactions de l'utilisateur en transformant les signaux de bas niveau, tels que les expressions faciales, la rotation de la tête, etc., en variables de haut niveau, telles que le degré d'impression de l'utilisateur ou l'engagement de l'utilisateur (voir paragraphe suivant); (2) un pour adapter les comportements de l'agent selon les réactions de l'utilisateur; (3) un pour générer le comportement de l'agent.

L'objectif du module d'adaptation (2) est de donner à l'agent la capacité de faire face aux réactions de l'utilisateur en temps réel pendant l'interaction, et d'adapter son comportement réalisé en fonction de ses propres intentions (c'est-à-dire gérer les impressions de C&C) et des impressions formées par l'utilisateur. Le but de l'adaptation du comportement est d'avoir une meilleure gestion des impressions et d'être en mesure de maintenir les impressions souhaitées.

Cette architecture est décrite dans le chapitre 7.

Étape 4 : Cas d'utilisation

Une fois que l'architecture générale de gestion d'impressions a été réalisée, nous avons conçu deux cas d'utilisation où nous avons appliqué ce système à un scénario d'une vraie interaction, afin d'évaluer l'impact du modèle de gestion d'impressions sur l'interaction utilisateur-agent.

Les Premières Impressions

La formation d'une impression sur un inconnu peut être décrite en trois étapes. Tout d'abord, nous *percevons* la nouvelle personne rencontrée, souvent visuellement, car la vision est le canal sensoriel le plus rapide. Les gens perçoivent et recueillent immédiatement de l'information sur les caractéristiques invariantes comme l'âge, le sexe, l'origine ethnique, et les caractéristiques variables comme le visage, les gestes, la posture du corps et le regard. Après avoir acquis cette première information, les gens font des *inférences* sur l'autre, par exemple sur sa personnalité. Enfin, la nouvelle personne est *catégorisée* dans un certain groupe social. Les stéréotypes influencent souvent les processus de formation d'impressions (Fiske et al., 2002). A partir des années 40, plusieurs théories de la cognition sociale ont fourni des explications sur le processus de collecte et de traitement de l'information sur l'autre. Deux approches ont émergé de cet axe de recherche. Une approche a considéré les impressions dans leur globalité (en suivant les principes de la psychologie de la Gestalt), comme le résultat de la relation entre les traits individuels. Ceci a d'abord été proposé par Asch (1946). La deuxième approche a appliqué des règles mathématiques pour calculer l'impression finale à partir des valeurs des traits individuels. Elle a été proposée par Anderson (1968). Plus récemment, d'autres approches ont intégré des éléments cognitifs et motivationnels dans le processus de formation de l'impression. Le plus pertinent est le modèle de Fiske and Neuberg (1990) qui souligne le rôle des stéréotypes, de la motivation personnelle et des ressources d'attention dans la formation des impressions.

En plus de se forger des impressions sur les autres, les gens essaient souvent de contrôler les impressions que les autres se forgent sur eux. La gestion d'impressions "décrit les efforts déployés par un acteur pour créer, maintenir, protéger ou autrement modifier une image détenue par un public cible" (Bozeman and Kacmar, 1997). Leary and Kowalski (1990) ont passé en revue les principales théories sur les facteurs qui influencent la gestion d'impressions, en particulier le point de vue social de Goffman et al. (1978), l'approche psychologique de Jones and Pittman (1982) et d'autres approches telles que celle de Schlenker (1980). À partir de cet examen, ils ont identifié deux composantes principales qui sous-tendent la gestion d'impressions et ils ont proposé un modèle pour expliquer leur rôle dans ce processus.

Le premier facteur qui affecte la gestion des impressions concerne la motivation de créer une impression particulière dans les autres, mais cela n'implique pas nécessairement que cette impression soit réalisée. Dans le modèle à deux composantes de Leary and Kowalski (1990), trois facteurs principaux sont identifiés qui affectent la motivation de l'impression: l'*importance des objectifs des impressions*, le *valeur des objectifs souhaités* et le *disparité entre l'image souhaitée et l'image actuelle*. Dans le cadre de notre travail, ce dernier facteur représente la principale motivation de notre ACA: les impressions de l'utilisateur étaient constamment surveillées et lorsque ces impressions ne correspondaient pas au but de l'agent, il les gérait en adaptant son comportement afin d'obtenir l'impression souhaitée chez l'utilisateur.

Le deuxième facteur qui influe sur la gestion des impressions concerne le type d'impression que l'on cherche à transmettre et la manière de la réaliser, c'est-à-dire les moyens à utiliser. Nous pouvons contrôler notre impression en termes de traits de personnalité, d'attitudes, de rôles, de croyances, etc. Pour ce faire, on peut utiliser l'auto-description, le comportement non verbal, l'apparence physique, l'association avec des groupes sociaux. Dans le modèle à deux composants de Leary and Kowalski (1990), cinq facteurs principaux sont

identifiés qui affectent la construction des impressions. Parmi ceux-ci, les gens peuvent choisir leurs caractéristiques qui correspondent aux préférences des autres. Cette stratégie n'est pas nécessairement trompeuse, par exemple les gens peuvent essayer d'omettre des informations qui ne correspondent pas aux valeurs de la cible (Leary and Lamphere, 1988). Ce facteur était important dans notre contexte puisque l'objectif de notre agent était d'adapter son comportement en fonction des préférences de l'utilisateur. Il avait un ensemble de comportements possibles à sélectionner, mais le choix final ne tenait compte que de ceux qui avaient été jugés positivement par l'utilisateur.

Dans le contexte de cette thèse, nous nous sommes intéressés au rôle du comportement non verbal dans la gestion d'impressions. Relativement peu d'œuvres ont montré comment les gens gèrent leur comportement non verbal afin de contrôler l'impression à donner aux autres. Par exemple, Rosenfeld (1966) a mené une expérience où les participants ont interagi en dyades sous 2 conditions. Dans la condition de demande d'approbation, on demandait à l'une des deux personnes de la dyade d'obtenir l'approbation de l'autre personne, tandis que dans la condition d'évitement de l'approbation, on lui demandait d'éviter l'approbation de l'autre personne. Les participants dans la condition de demande d'approbation ont produit plus de sourires, de hochements de tête, de gesticulations et de réactions verbales que les participants dans la condition d'évitement de l'approbation. De plus, l'approbation de l'autre membre de la dyade était positivement corrélée aux hochements de tête, à la réceptivité verbale et négativement corrélée aux adaptateurs et aux autoréférences.

La majorité des études expérimentales sur la gestion d'impressions ont révélé que l'auto-présentation concernait l'utilisation du comportement verbal plutôt que le comportement non verbal (e.g., Peeters and Lievens (2006)).

Dans notre contexte, puisque nous avons travaillé avec des ACA, nous avons pu contrôler et manipuler leur comportement non verbal. Notre objectif était d'étudier comment la gestion du comportement non verbal de l'agent peut influer sur la formation de l'impression de l'utilisateur sur l'agent.

Chaleur et Compétence

Lorsque nous rencontrons de nouvelles personnes, nous recueillons rapidement des informations sur leur intention vers nous, ainsi que sur leur capacité à atteindre cette intention. Nous utilisons les mêmes critères pour former des stéréotypes au sujet des groupes sociaux. Ces deux grandes dimensions ont été appelées avec des étiquettes différentes, qui se chevauchent dans leur signification. Nous avons choisi d'utiliser les termes chaleur et compétence: la première comprend des traits comme l'amabilité, la loyauté, la sociabilité; le deuxième comprend des traits comme l'intelligence, l'agence et l'efficacité. La chaleur et la compétence sont au cœur de la perception interpersonnelle et intergroupe et elles suscitent des résultats émotionnels et comportementaux uniques. Plusieurs expériences supportent la primauté des jugements liés à la chaleur sur la compétence. Il n'y a pas d'accord sur le type de relation entre les deux dimensions : elles sont indépendantes selon certains cadres, alors qu'en général, elles sont corrélées positivement dans un contexte à cible unique, et négativement dans un contexte de comparaison à deux cibles. Nos impressions sur la chaleur et la compétence des autres ne sont pas seulement obtenues par l'observation ou la description de comportements manifestes (comme les actions), mais elles peuvent aussi être suscitées par des indices non verbaux particuliers, comme des postures ouvertes ou fermées, des types de gestes et le sourire. Une étude intéressante sur l'effet des gestes de

la main sur la perception sociale (Maricchiolo et al., 2009) a montré des effets significatifs du type de gestes de la main sur la perception des compétences. Dans leur expérience, les chercheurs ont créé 5 vidéos différentes où un acteur jouait le rôle d'un délégué universitaire discutant de la décision du Conseil universitaire d'augmenter les frais de scolarité. Les vidéos ne différaient que par le type de gestes accomplis par l'acteur : idéationnels (c'est-à-dire des gestes liés au contenu sémantique du discours), rythmiques (gestes rythmiques, liés à la structure et au rythme du discours), adaptateurs d'objets (mouvement des mains au contact des objets), auto-adaptateurs (mouvement des mains au contact des parties du propre corps) et absence de geste. Les auteurs ont ensuite demandé aux participants d'évaluer l'approche communicative du conférencier, son style d'orateur, la force de persuasion du message, leurs attitude (favorable ou défavorable) face à l'augmentation des frais, et leur intention de voter à ce sujet. Les gestes idéationnels et les adaptateurs d'objets ont donné lieu à des jugements de compétence plus élevés, comparativement à l'absence de gestes, tandis que les auto-adaptateurs ont donné lieu à une perception de compétence plus basse. Aucun effet significatif des gestes de la main n'a été trouvé pour la chaleur.

Cette dernière étude est l'un des rares travaux qui a donné un aperçu du rôle des gestes de communication dans la transmission des différentes impressions de C&C. Dans cette thèse, nous avons voulu approfondir non seulement les différents rôles du type de gestes (idéationnels, bâtons, adaptateurs) mais aussi le rôle des positions de repos des bras. Peu de recherches ont étudié cette question, en particulier l'association entre les poses de repos et la dominance, mais pas explicitement avec les impressions C&C. Les travaux présentés au chapitre 5 ont été réalisés à cette fin. Nous avons étudié l'association entre les impressions C&C et le comportement non verbal comme le type de gestes, les positions de repos des bras, les sourires et les rotations de la tête.

Etat de l'Art

Bien que plusieurs études aient porté sur le rôle du comportement non verbal dans la formation d'impressions, peu d'entre elles se sont concentrées explicitement sur l'impact des comportements sur les dimensions de la chaleur et de la compétence. Les études portant sur l'effet de l'apparence de l'agent sur les impressions des utilisateurs montrent que la gestion de l'apparence et de la voix ne suffit pas pour obtenir des impressions cohérentes de la chaleur et de la compétence de l'agent et nous encouragent à prendre en compte le rôle des comportements non verbaux. Lorsque l'on examine les études sur la chaleur et la compétence des agents virtuels, leurs résultats sont conformes aux phénomènes que nous avons décrits au chapitre 3. Il semble que les mêmes schémas se produisent lorsque les gens jugent les agents virtuels. En particulier, le soutien pour *l'effet de halo* a été trouvé par Niewiadomski et al. (2010) et Nguyen et al. (2015), une *primauté des jugement de chaleur* a été trouvé par Niewiadomski et al. (2010) et les résultats du Bergmann et al. (2012) reflètent la présence d'une *diagnostique asymétrique* de chaleur et compétence. En ce qui concerne ces études, nous visons toujours à étudier la nature des impressions de chaleur et de compétence dans l'interaction humain-agent, en mettant l'accent sur les relations entre les deux dimensions. En ce qui concerne **?**, nous avons considéré plus de comportements que de simples gestes liés au discours, tels que les expressions faciales, l'inclinaison du tronc et les poses de la tête, et nous avons mis en place un modèle de gestion d'impressions. Pour trouver ce que sont ces comportements, nous avons proposé une

méthodologie qui a utilisé des vidéos d'interactions naturelles au lieu de vidéos d'acteurs (voir Chapitre 5) avec des annotations à la fois discrètes et continues.

Nous avons ensuite passé en revue les principales études menées dans des espaces publics comme les musées des sciences. Aucun d'entre elles ne traitait de la gestion des impressions et ne proposait un système intégrant l'évaluation des impressions des utilisateurs, notamment la détection des indices multimodaux des utilisateurs pour la mise en œuvre des mécanismes d'adaptation du comportement de l'agent. De plus, ils se concentrent principalement sur le dialogue de l'agent, alors que dans cette thèse nous nous intéressons à la gestion en temps réel du comportement non verbal de l'agent.

Dans le premier cas d'utilisation de cette thèse, nous avons pris en compte l'engagement de l'utilisateur pendant l'interaction. Nous avons passé en revue plusieurs travaux axés sur la promotion de l'engagement de l'utilisateur en utilisant différentes stratégies et différentes méthodes de détection de l'utilisateur. La méthode que nous avons utilisée pour détecter l'engagement de l'utilisateur, décrite au chapitre 8, était similaire à certaines d'entre elles mais prenait en compte à la fois les expressions du visage et la rotation de la tête et du tronc de l'utilisateur.

Chaleur et Compétence dans l'interaction humain-humain

La première étape de l'approche suivie dans cette thèse a consisté à étudier la perception de C&C en analysant les interactions naturelles entre les humains. L'objectif était d'identifier les comportements non verbaux qui peuvent provoquer différents degrés de C&C, puisque dans la littérature nous avons trouvé relativement peu d'informations à leur sujet (voir Section 3.5). Nous nous sommes concentrés sur le rôle du type de gestes, les positions de repos des bras (c.-à-d. la position des bras lorsqu'ils ne font aucun geste), les mouvements de la tête et les sourires.

Pour l'analyse de l'interaction humain-humain, nous avons cherché un corpus à analyser dont la mise en place était similaire à une interaction typique entre un humain et un agent virtuel. En particulier, nous avons cherché un corpus correspondant à 4 critères : nous aimerions analyser les interactions dyadiques, où les participants se comportaient de manière naturelle et spontanée, où des connaissances étaient partagées entre les participants, et où des enregistrements du comportement corporel complet étaient disponibles.

Nous avons calculé l'association entre des annotations discrètes de comportements non verbaux (type de gestes, positions de repos des bras, mouvements de tête, sourires) et des annotations de la chaleur et de la compétence de l'expert perçues (converties de continues en deux niveaux discrets décrivant l'augmentation et la diminution).

Les données continues ont été prétraitées afin de tenir compte du délai de réaction, de réduire le bruit et d'envisager un accord relatif entre les annotateurs plutôt qu'absolu. Seules les fenêtres temporelles où les annotateurs étaient d'accord sur le type de variation de chaleur (ou de compétence) exprimée par l'expert ont été conservées.

Les résultats ont montré le rôle important du sourire. Le sourire était associé à des jugements d'augmentation de la chaleur et de diminution de la compétence. Ceci est en ligne avec les résultats précédents (Bayes, 1972; Cuddy et al., 2011), et suggère la preuve d'un effet de compensation entre les deux dimensions fondamentales de la cognition sociale. Le sourire a également eu un impact important sur l'association de certains types de gestes et de positions de repos des bras avec les jugements de C&C. Par exemple, lorsque les experts croisaient les bras, les jugements de compétence diminuaient, mais la direction

de cette association était inversée lorsque la même position de repos se produisait avec un sourire. Nous avons observé un effet similaire entre les bras croisés et la chaleur.

La relation entre les adaptateurs effectués en souriant avec des jugements de plus élevés de compétence semble être en contraste avec les résultats d'autres études. Les auto-adaptateurs ont souvent été associés à des manifestations de stress et d'anxiété (**?**), qui se traduisent par un faible niveau de compétence perçue. Cependant, notre résultat pourrait s'expliquer par le fait que les sourires adoucissaient la relation des auto-adaptateurs avec le stress et rendaient plus évidente, pour l'observateur, la perception de la compétence. Les positions de repos des bras ont contribué à diminuer les jugements pour les deux dimensions. C'est surprenant étant donné que dans les travaux précédents, on n'a trouvé aucun lien entre ces comportements et les premières impressions de C&C.

Pour la majorité des indices non verbaux observés (à l'exception du sourire), nous avons trouvé des preuves à l'appui de l'effet de halo. Plus précisément, les niveaux C&C allaient dans la même direction. Les résultats ont également confirmé la primauté de la chaleur sur la compétence en termes d'ampleur de l'effet.

En ce qui concerne les mouvements de la tête, nous avons trouvé des tendances prometteuses (entre les hochements de tête et le niveau de compétence et entre les inclinaisons et la chaleur) mais sans atteindre de signification statistique.

Perception de Chaleur et de Compétence dans un Agent Virtuel

A partir des résultats obtenus par les analyses décrites au chapitre 5, la deuxième étape de notre approche a consisté à comprendre si les mêmes processus caractérisant la perception sociale dans les interactions humain-humain s'appliquent à la perception d'un ACA, en particulier s'il est possible pour un ACA d'exprimer différents degrés de C&C par son comportement non verbal.

De plus, nous avons examiné le rôle des attentes des gens à l'égard des ACA. Burgoon (1993) a déclaré que les gens ont des attentes à l'égard du comportement des autres au cours d'une conversation, qui sont principalement fondées sur les normes sociales et les caractéristiques spécifiques des cibles. Ces attentes peuvent être confirmées ou violées pendant l'interaction. La théorie de Burgoon sur la violation des attentes soutient que les violations de ces attentes entraînent généralement des résultats plus extrêmes que les confirmations de celles-ci.

Burgoon et al. (2016) ont étudié la validité de leur théorie dans le cas de l'interaction humain-agent : il semble que nous ayons des attentes concernant le comportement des ACA, et que ces attentes puissent être violées. Leurs conclusions nous ont encouragés à étudier comment les attentes pourraient influer sur la perception de C&C d'un ACA.

L'étude réalisée pendant cette étape visait à répondre aux questions de recherche suivantes :

- **(Q1a)** *Un ACA est-il perçu différemment en termes de C&C selon les comportements non verbaux qu'il réalise ?*
- **(Q1b)** *Si oui, quels sont les comportements non verbaux (ou combinaisons de ces comportements) qui lui permettent d'être mieux perçu en termes de C&C ?*
- **(Q2)** Nos attentes et a priori sur les ACA influencent-ils nos impressions suivantes sur l'ACA ?

Pour répondre à ces questions, nous avons conçu une étude perceptive où nous avons manipulé certains comportements non verbaux dans un agent virtuel. Les choix concernant les signaux et la conception de l'étude résultent du compromis entre le désir d'étudier tous les signaux qui nous intéressent et la nécessité de limiter la complexité de la conception de l'étude.

Les résultats de cette étude perceptive ont montré l'influence du type de geste sur la perception de C&C; en particulier, lorsque l'agent faisait des gestes idéationnels (liés à ce dont il parlait), il était perçu comme plus chaleureux et plus compétent que lorsqu'il faisait des gestes de battement dont les formes n'étaient pas liées au contenu du discours. L'utilisation de ces gestes peut refléter la motivation de l'agent à aider l'utilisateur à mieux comprendre de quoi il parlait et, en même temps, sa connaissance du sujet.

En ce qui concerne la chaleur, les gestes idéationnels ont eu un effet positif sur la perception de cette dimension seulement quand ils étaient réalisés à haute fréquence.

En ce qui concerne l'hypothèse d'un effet des attentes sur les jugements sur l'agent, lorsque l'agent était présenté comme intelligent et autonome, cela affectait l'évaluation des participants par rapport à quand l'agent était présenté comme une marionnette contrôlée par un humain. Ce résultat semble confirmer le rôle des attentes des gens sur la formation d'impressions et souligne l'importance de tenir compte des a priori des participants au sujet des agents virtuels.

Aucun effet de la fréquence des sourires n'a été constaté. Ceci pourrait s'expliquer par le fait que la présence d'un seul sourire était déjà suffisante pour donner une impression et que cela ne variait pas si on augmentait la fréquence de ce signal.

Enfin, nous n'avons trouvé aucun effet des positions de repos des bras. En fait, pour des raisons d'animation, l'agent n'a effectué une position de repos spécifique qu'au début et à la fin de la vidéo, tandis que pendant le reste de la vidéo, la position de repos consistait à mettre les bras le long de son corps. Il est probable que la présence des différentes postures de repos à l'étude soit trop subtile pour que l'on puisse percevoir une différence entre les conditions.

Par contre, selon les commentaires des participants, l'expérience globale a été jugée assez longue et épuisante, même si la durée totale de l'expérience n'a pas dépassé 20 minutes.

Il est intéressant de comparer ces résultats avec ceux obtenus dans l'étude de l'interaction humain-humain décrite dans le paragraphe précédent. En particulier, comme pour les résultats précédents, nous avons trouvé un *effet de halo* des gestes sur les jugements de C&C, car ils allaient dans la même direction pour la chaleur et la compétence. En plus de l'étude précédente, nous avons trouvé ici un effet des gestes idéationnels sur la perception de la compétence, alors qu'auparavant ce comportement non verbal n'était pas associé de façon significative à cette dimension. Contrairement aux résultats de l'analyse de l'interaction humain-humain, nous n'avons trouvé aucun effet du sourire sur la perception de la chaleur et de la compétence de l'agent par l'utilisateur, ni de l'*effet de compensation* que nous avons trouvé dans l'étude précédente.

Architecture pour la gestion d'impressions de l'agent

Le but principal de cette thèse était de construire un modèle computationnel pour un agent conversationnel animé capable de gérer ses impressions de C&C envers l'utilisateur. L'architecture que nous avons conçue pour doter l'ACA de la capacité d'adapter son comportement aux réactions des utilisateurs est suffisamment générale pour permettre la per-

sonnalisation de ses différents modules en fonction des différents contextes et objectifs de l'agent. Deux exemples d'application de cette architecture sont présentés dans les chapitres 8 et 9, où l'architecture a été personnalisée afin d'adapter le comportement de l'agent en fonction de l'engagement de l'utilisateur et de ses impressions à son égard, respectivement.

Les travaux présentés dans ce chapitre ont été réalisés en collaboration avec l'étudiant de Master Paul Lerner[1] et le Professeur Maurizio Mancini[2].

L'objectif principal du modèle était de gérer le comportement non verbal de l'agent pour obtenir différentes impressions de C&C en fonction des réactions des utilisateurs. Pour ce faire, 3 modules principaux ont été impliqués :

1. Le *Module d'analyse de l'utilisateur* pour détecter et interpréter les réactions de l'utilisateur;
2. Le *Module de gestion d'impressions* pour sélectionner l'impression à obtenir, à travers les intentions communicatives de l'agent ;
3. Le *Module de génération du comportement de l'agent* pour l'animation de l'agent.

Nous avons inclus un planificateur de dialogue dans le *Module de gestion d'impressions*, même si nous avons essayé de garder le dialogue aussi basique que possible, car au-delà de nos intérêts de recherche.

Comme nous l'avons dit précédemment, le but de l'agent était d'adapter ses comportements à chaque participant. Cela impliquait de donner à l'agent la capacité d'apprendre en temps réel quel était le meilleur comportement à adopter, en fonction de son objectif (par ex. obtenir une impression liée à la chaleur) et des réactions de l'utilisateur (par ex. l'impression de l'utilisateur concernant la chaleur de l'agent).

Un algorithme d'apprentissage par renforcement nous a semblé la meilleure approche pour nos besoins, car il n'exige pas une connaissance préalable de l'environnement et a pour but de maximiser une récompense au lieu de découvrir des structures cachées dans les données. Le cadrage typique de l'apprentissage par renforcement inclut une boucle où un agent choisit une *action* dans un *environnement*, et ces actions sont associées à une *récompense* et à une représentation d'un *état*, qui sont renvoyés à l'agent. Cela correspond bien à notre cadre général où l'agent adopterait des comportements dans l'environnement des états d'interaction, recevrait une récompense de la réaction des utilisateurs et l'utiliserait pour adapter son comportement. Comme l'apprentissage par renforcement permet à l'agent d' "apprendre de l'interaction" (Sutton and Barto, 2018), il doit relever le défi de trouver un équilibre entre exploration et exploitation. Autrement dit, l'agent doit exploiter ce qu'il a déjà vécu afin d'obtenir une récompense, mais il doit aussi explorer afin de faire de meilleures sélections d'actions à l'avenir. Le dilemme est que ni l'exploration ni l'exploitation ne peuvent être poursuivies exclusivement sans échouer à la tâche. L'agent doit essayer une variété d'actions et favoriser progressivement celles qui lui semblent les meilleures.

Nous avons mis en place des modules pour capturer le comportement de l'utilisateur (parole, regard, expressions faciales, orientation de la tête et du tronc), l'analyser/interpréter (par exemple, détecter les impressions de l'utilisateur sur l'agent) et décider ce que l'ACA doit dire et comment (par exemple, les comportements non verbaux accompagnant le discours).

On peut distinguer 3 parties principales dans notre modèle :

[1] UFR de Mathématiques et Informatique, Université Paris Descartes

[2] School of Computer Science and Information Technology, University College Cork

1. *Analyse de l'utilisateur.* Nous avons exploité la plateforme EyesWeb (Camurri et al., 2004) pour extraire en temps réel : (1) les signaux non verbaux de l'utilisateur (par exemple, la rotation de la tête et du tronc) à partir des données du squelette capturées par une Kinect ; (2) les unités d'action du visage de l'utilisateur, en exécutant le logiciel OpenFace (Baltrušaitis et al., 2016) ; (3) le regard de l'utilisateur grâce à l'eye tracker Tobii ; (4) le discours de l'utilisateur en exécutant la plateforme Microsoft Speech Platform https://www.microsoft.com/en-us/download/details.aspx?id=27225.

   Ces signaux de bas niveau sont traités par EyesWeb et d'autres outils externes, tels que des modèles déjà entrainés d'apprentissage automatique, pour extraire des caractéristiques de haut niveau sur l'utilisateur.

2. *Gestion d'impressions.* C'était le module de prise de décision du système, où l'information de l'utilisateur est exploitée par un algorithme d'apprentissage par renforcement et la parole de l'utilisateur peut être traitée par des outils de traitement du langage naturel et envoyée à un planificateur de dialogue. Le résultat du module inclut le comportement verbal et non verbal provoquant différents niveaux de C&C.

3. *Animation de l'Agent.* La génération du comportement de l'agent est réalisée par VIB/Greta, une plateforme supportant la création d'agents conversationnels socio-émotionnels (Pecune et al., 2014). VIB/Greta génère l'animation de l'ACA composée de gestes, d'expressions faciales et du regard, en synchronie avec la parole.

Cas d'utilisation n.1: adaptation des comportements de l'agent selon l'engagement de l'utilisateur

Dans notre première application de l'architecture du système de gestion des impressions d'ACA en temps réel, le cas d'utilisation était un agent jouant le rôle d'un guide virtuel de musée. Notre but était de gérer les dimensions C&C afin d'obtenir un ACA engageant, en suivant l'idée qu'un agent plus engageant est susceptible de former une impression positive et d'être accepté par l'utilisateur, favorisant ainsi des interactions ultérieures (Bergmann et al., 2012; Cafaro et al., 2017). D'autres auteurs se sont concentrés sur différentes stratégies pour améliorer l'engagement de l'utilisateur, par exemple sur les backchannels, les stratégies de politesse ou l'alignement verbal (voir Section 4.4). Dans cette thèse, puisque nous nous sommes concentrés sur les impressions de C&C, nous nous sommes intéressés au l'impact de ces impressions sur l'engagement de l'utilisateur, en particulier si l'adaptation des impressions de C&C de l'agent pourrait affecter l'engagement de l'utilisateur.

Suivant ce raisonnement, nous nous sommes concentrés sur deux grandes questions de recherche :

- (Q1) *Existe-t-il une relation entre les impressions de C&C de l'agent et l'engagement de l'utilisateur pendant l'interaction avec un ACA ?*

- (Q2) *Est-il possible d'améliorer l'engagement de l'utilisateur en gérant le degré de C&C de l'agent ?*

Afin de répondre à ces questions, nous nous sommes concentrés sur les effets des stratégies d'auto-présentation qui pourraient être réalisées par l'agent afin de gérer ses impressions

de C&C. Par exemple, l'agent peut décider de se présenter comme un guide chaleureux, ou de mettre en valeur son niveau de compétence en diminuant sa chaleur. Au début de l'interaction, l'agent n'avait aucune information sur les effets de sa stratégie sur la perception de l'utilisateur. Il pourrait utiliser le niveau d'engagement de l'utilisateur comme mesure pour évaluer quelle stratégie a le mieux fonctionné, c'est-à-dire quelle stratégie a accru l'engagement de l'utilisateur.

Nous avons personnalisé le module *Analyse de l'utilisateur* afin de calculer l'engagement de l'utilisateur à partir de signaux de bas niveau et l'utiliser comme récompense pour l'algorithme d'apprentissage par renforcement. Le module *Gestion d'impressions* a également été adapté en incluant un planificateur d'intention d'auto-présentation.

Jones and Pittman (1982) ont affirmé que les gens peuvent utiliser différentes techniques comportementales verbales et non verbales pour créer les impressions qu'ils souhaitent chez leur interlocuteur. Les auteurs proposent une taxonomie de ces techniques, qu'ils appellent stratégies d'auto-présentation. Nous illustrons ici 4 de leurs stratégies qui peuvent être associées à différents niveaux de C&C. Nous n'avons pas considéré la 5ème stratégie de la taxonomie, appelée *Exemplification*. Cette stratégie est utilisée lorsque les gens veulent être perçus comme étant dévoués et obtenir l'attribution du dévouement des autres, donc elle n'est liée ni à la chaleur ni à la compétence. Concernant les 4 autres stratégies, deux d'entre elles se concentrent sur une dimension à la fois, les deux autres se concentrent sur les deux dimensions en leur donnant des valeurs opposées :

- *Ingratiation* : son but est d'amener l'autre personne à vous apprécier et à lui attribuer des qualités interpersonnelles positives (par exemple, chaleur et gentillesse). Dans notre cas, l'agent qui a choisi cette stratégie avait pour but de susciter des impressions de chaleur chez l'utilisateur, sans tenir compte de son niveau de compétence.

- *Supplication* : elle se produit quand les individus présentent leurs faiblesses ou leurs déficiences pour recevoir la compassion et l'aide des autres. Dans notre cas, l'agent qui a choisi cette stratégie avait pour but de susciter des impressions de chaleur élevée et de faible compétence.

- *Self-promotion* : elle se produit quand les individus attirent l'attention sur leurs réalisations pour être perçus comme capables par les observateurs. Dans notre cas, l'agent qui a choisi cette stratégie avait pour but de susciter des impressions de haute compétence, sans tenir compte de son niveau de chaleur.

- *Intimidation* : elle est définie comme la tentative de projeter son propre pouvoir ou sa propre capacité à punir pour être considéré comme dangereux et puissant. Dans le contexte de notre recherche, nous avons interprété cette stratégie d'une manière plus souple, comme le but de susciter des impressions de faible chaleur et de compétence élevée.

Dans notre cas d'utilisation, pour chaque tour de parole, l'agent jouait une de ces 4 techniques d'auto-présentation.

Ces techniques ont été réalisées par l'ACA à travers son comportement verbal et non verbal. Le comportement verbal caractérisant les différentes stratégies s'inspire des travaux de Pennebaker (2011) et Callejas et al. (2014). Selon leurs conclusions, nous avons manipulé l'utilisation des pronoms *toi* et *moi*, le niveau de formalité de la langue et la longueur des phrases. Par exemple, les phrases visant à susciter une chaleur élevée contenaient plus de pronoms, moins de synonymes, un langage plus informel, de sorte que

les phrases étaient plus décontractées et donnaient l'impression d'être moins méditées ; plus de verbes que de noms et un contenu positif était prédominant. Les phrases visant à susciter une faible chaleur contenaient plus de négations, des phrases plus longues, un langage formel et ne faisaient pas référence à l'orateur. Les phrases visant à obtenir une compétence élevée contenaient des taux élevés des mots “ nous " et “ vous ", et le mot “ je " à faible taux.

Le choix du comportement non verbal de l'agent était basé sur nos études précédentes décrites au chapitre 5 et 6. En particulier, nous avons manipulé le type de gestes et le type de positions de repos des bras et les sourires.

Ainsi, par exemple, si la stratégie d'auto-présentation de l'agent actuel était *Supplication* et que l'acte de dialogue suivant était d'introduire un sujet, alors l'agent dirait : “Je pense que pendant que tu joues, il y a des capteurs qui mesurent des tonnes de choses !” accompagné d'un bâton et d'un sourire. Inversement, si la stratégie d'auto-présentation de l'agent actuel était *Intimidation* et que l'acte de dialogue suivant était le même, alors l'agent dirait : “Pendant que tu joues aux jeux vidéo, plusieurs capteurs mesurent tes signaux physiologiques”, sans sourire ni gestes.

Nous avons personnalisé l'architecture générale de la gestion d'impressions de l'agent en temps réel afin d'adapter les intentions de présentation de l'utilisateur selon l'engagement de l'utilisateur. Nous avons construit une architecture qui prend en entrée les unités d'action du visage des participants et la rotation du tronc et de la tête, les utilise pour calculer l'engagement global de l'utilisateur et envoie celui-ci au module de gestion d'impressions de l'agent. Grâce à un algorithme de multi-armed bandit qui prenait l'engagement de l'utilisateur comme récompense, l'agent pouvait sélectionner l'intention d'auto-présentation maximisant ainsi l'engagement de l'utilisateur. Afin d'évaluer le système, nous avons conçu un scénario d'interaction où l'agent jouait le rôle de guide de musée. Dans l'expérience, nous avons manipulé la façon dont l'agent choisissait son intention d'auto-présentation à chaque tour de parole. Il pourrait adapter son comportement en utilisant l'algorithme d'apprentissage par renforcement, ou le choisir au hasard, ou utiliser la même intention d'auto-présentation pendant toute la durée de l'interaction.

L'agent qui a adapté son comportement pour maximiser l'engagement de l'utilisateur a été perçu comme plus chaleureux par les participants. De plus, nous avons trouvé un lien entre l'adaptation de l'agent, l'engagement de l'utilisateur et les impressions de chaleur : plus l'agent adapte ses comportements, plus l'utilisateur est engagé et plus il perçoit l'agent comme chaleureux.

Cas d'utilisation n.2: adaptation des comportements de l'agent selon les impressions de l'utilisateur

Dans notre deuxième application de l'architecture du système pour la gestion des impressions d'ACA en temps réel, en utilisant le même scénario que dans le cas d'utilisation précédent (avec quelques petits changements dans la configuration), nous avons voulu tester un modèle de détection développé par Chen Wang, Guillaume Chanel et Thierry Pun de l'Université de Genève, les partenaires du projet IMPRESSION. Ce modèle a permis de détecter l'impression de chaleur ou de compétence que l'utilisateur se forse sur l'agent en analysant l'activité des unités d'action de l'utilisateur. Rappelons que les techniques d'auto-présentation véhiculant différents niveaux de C&C n'ont pas été complètement validées dans l'expérience précédente (aucune différence significative entre les dif-

férentes stratégies n'a été observée pour les évaluations de compétence et une seule technique différait des autres en termes de scores de chaleur). En exploitant le modèle de détection développé par nos partenaires, notre objectif était que l'agent puisse gérer son comportement de manière à afficher l'impression de chaleur ou de compétence la plus appropriée.

Avec cette deuxième expérience, nous avons tenté de répondre aux questions de recherche suivantes :

- (Q1) *Est-il possible d'obtenir différentes impressions de C&C en adaptant le comportement de l'agent en fonction des impressions de l'utilisateur?*
- (Q2) *Est-il possible d'influencer la perception de l'interaction par l'utilisateur en maximisant la chaleur (ou la compétence) de l'agent pendant l'interaction ?*

Pour répondre à ces questions, nous avons personnalisé le *Module d'analyse de l'utilisateur* pour calculer les impressions de l'utilisateur à partir de signaux de bas niveau et les utiliser comme récompense pour l'algorithme d'apprentissage par renforcement. Le *Module de gestion d'impressions* a également été modifié en adaptant l'algorithme d'apprentissage par renforcement et en incluant un ensemble de comportements verbaux et non verbaux possibles à exécuter. Dans ce cas, nous n'avons pas créé d'intentions d'auto-présentation mais nous avons donné à l'agent un ensemble de comportements qu'il pouvait combiner comme il voulait (c'est-à-dire en exécutant l'algorithme d'apprentissage par renforcement). Nous avons mené une étude d'évaluation où nous avons comparé un agent qui adapte son niveau de chaleur ou de compétence à un agent non adaptatif. Les résultats ont montré que l'agent adaptatif a réussi à influencer les impressions de l'utilisateur sur sa compétence, tandis que les impressions a priori des utilisateurs ont affecté leurs impressions sur la chaleur de l'agent.

Contributions de cette Thèse

Première contribution: *Création d'un répertoire de comportements multimodaux suscitant des impressions de chaleur et de compétence.*

Durant la première phase de notre travail, notre but était de trouver des associations entre les comportements non verbaux associés à la chaleur et les impressions de compétence. Nous sommes partis de l'étude de la littérature sur les indices non verbaux de chaleur et de compétence et des études existantes qui incluaient ces dimensions dans l'interaction humain-agent. Après cette première étude, nous avons suivi une approche guidée par 2 motivations principales. Premièrement, nous n'avons pas trouvé une grande quantité d'information dans la littérature sur le comportement non verbal qui suscite des impressions de chaleur et de compétence. Deuxièmement, les quelques travaux concernant le comportement non verbal d'un agent virtuel suscitant des impressions de chaleur et de compétence s'appuyaient sur un corpus d'acteurs, alors que nous voulions recueillir des informations à partir de l'étude des interactions naturelles. C'est pourquoi nous avons annoté et analysé le corpus NoXi. Les annotations ajoutées au corpus sont disponibles sur https://noxi.aria-agent.eu/ et peuvent être utiles à d'autres chercheurs pour d'autres analyses ou pour produire d'autres annotations en suivant le même schéma d'annotation que nous avons utilisé. Nous avons contribué à donner un aperçu du rôle du type de gestes, des positions de repos des bras et du sourire dans la formation de ces impressions.

Nous avons trouvé des correspondances avec la littérature telles que la présence de l'effet de halo (Rosenberg et al., 1968) pour les gestes et l'effet de compensation (Yzerbyt et al., 2008) pour le sourire. Une autre contribution provient des résultats de l'étude perceptive présentée au chapitre 6, qui souligne le rôle des attentes dans le jugement des agents virtuels, conformément à la théorie de Burgoon et al. (2016).

Seconde contribution: Création d'un module d'adaptation du comportement pour la gestion des impressions de l'ACA.

Nous avons créé un module basé sur l'apprentissage par renforcement permettant à l'ACA d'adapter son comportement aux réactions des utilisateurs. Il permet de définir différentes récompenses pour les comportements utilisés par l'agent. Ce module permet d'apprendre en temps réel sans avoir une connaissance préalable des réactions de l'utilisateur à son comportement. Il permet de tester toutes les combinaisons possibles de comportements verbaux et non verbaux afin de trouver le meilleur pour produire une certaine impression sur l'utilisateur. Il peut être adapté aux différents objectifs de l'agent et permet de mieux comprendre le rôle des comportements non verbaux dans l'interaction humain-agent.

Troisième contribution: Création d'un ensemble de stratégies pour gérer les impressions de chaleur et de compétence dans un ACA.

En partant des résultats de l'analyse de l'interaction humain-humain, nous avons étudié le rôle des comportements multimodaux et des attentes dans le jugement des agents virtuels, afin de créer un ensemble de stratégies pour gérer les impressions de chaleur et de compétence dans un ACA. Ces stratégies s'inspirent de la taxonomie de Jones and Pittman (1982). Selon la stratégie choisie, l'agent adopte un comportement verbal et non verbal dans le but d'être perçu comme chaleureux, compétent, chaleureux et non compétent, ou froid et compétent. Ces stratégies ont été partiellement validées dans notre étude d'évaluation, notamment pour la dimension de la chaleur, et pourraient être mises en œuvre dans l'architecture générale de gestion d'impressions de l'agent.

Quatrième contribution: Mise en œuvre du module de gestion d'impressions dans une architecture système pour une interaction humain-agent en temps réel.

Nous avons intégré le module d'apprentissage par renforcement pour la gestion d'impressions de l'agent dans une architecture comprenant un module de détection et d'interprétation des réactions multimodales de l'utilisateur et un module de génération du comportement de l'agent. L'architecture est suffisamment générale pour permettre la personnalisation des différents modules en fonction des différents contextes et objectifs de l'agent. Avec ce travail, nous avons apporté une forte contribution en concevant un cadre interactif humain-agent qui peut être adapté et exploité dans d'autres projets. Par exemple, la possibilité de mettre en place des modules de détection des données physiologiques de l'utilisateur permettrait de mieux comprendre l'impact du comportement de l'agent sur l'état affectif de l'utilisateur. Ce travail offre un impact potentiel important sur de nombreuses applications telles que les assistants web et le déploiement d'agents en situation réelle (par exemple dans les gares ou les musées).

Cinquième contribution: Etude de l'efficacité d'un agent adaptatif et de la relation entre

l'adaptation, l'engagement et les impressions de l'agent, pendant l'interaction humain-agent.

Nous avons conçu un scénario où l'agent virtuel joue le rôle de guide de musée virtuel et personnalise l'architecture générale de gestion des impressions dans 2 applications différentes. Dans le premier, l'agent a adapté ses stratégies d'auto-présentation pour être perçu plus ou moins chaleureux ou compétent dans le but de maximiser l'engagement de l'utilisateur. Dans le second cas, il a adapté son comportement afin de maximiser les impressions de l'utilisateur sur son niveau de chaleur ou de compétence. Nous avons conçu et mené une étude d'évaluation pour chaque scénario afin de valider l'efficacité des capacités de gestion des impressions de l'agent. En particulier, nous voulions vérifier que les utilisateurs préféraient un agent possédant des compétences en gestion des impressions à un agent qui ne gérait pas les impressions des utilisateurs. Les études expérimentales menées à la Cité des sciences et de l'industrie ont été cruciales pour comprendre ce que les gens attendaient vraiment des ACA et ce qu'il faudrait améliorer dans notre modèle. Quelques résultats intéressants sont ressortis de nos études, montrant une certaine efficacité du modèle.

Mots-clefs : <agents conversationnels animés, interaction humain-agent, premières impressions, chaleur, compétence, comportement non verbal, >.

Abstract

LIKE in human-human interactions, the first moments of an interaction with a virtual character are critical since users form impressions about them, which can affect the rest of the interaction, in terms of engagement and willingness to continue it. In this Thesis we present a computational model for managing user's impression of agent's warmth and competence, the two fundamentals dimensions of social cognition. The goal of the agent is to adapt its non-verbal behaviour in real-time during an interaction, according to user's non-verbal reactions that are linked to his/her perceived impression of the agent. The methodology followed in this Thesis starts from the analysis of a corpus of human-human interactions in order to identify a set of non-verbal behaviours eliciting different degrees of warmth and competence. A perceptual study has then been conducted in order to investigate how these behaviours are perceived when performed by a virtual agent. Starting from the results of these first studies, a reinforcement learning algorithm has been developed to allow the agent to learn in real-time the behaviours which give the best impression to the user, according to its goal. User's non-verbal reactions (computed from low-level signals such as facial action units) are used as a reward for the reinforcement learning algorithm. We have personalized the computational model in order to adapt the agent's behaviours with the goal of maximizing (1) user's engagement and (2) user's impressions of agent's warmth and competence. Two use cases have been conducted at the "Cité des sciences et de l'industrie" in order to evaluate the impact of an adapting agent on user's impressions and perception of the interaction, compared to a non-adapting agent. In the first experiment the agent adapted its self-presentational strategies in order to maximise user's engagement. In the second experiment the agent learned the best combinations of non-verbal behaviours (e.g., gestures, arms rest poses, smiling) to display in order to maximise user's impressions of its warmth and competence.

Keywords: <human-agent interaction, first impressions, warmth, competence, non-verbal behaviour>.

Remerciements

Je voudrais tout d'abord remercier ma directrice de thèse, Catherine, qui m'a accueillie dans son équipe sans hésiter et qui m'a guidée tout au long de mon parcours. Merci pour ta gentillesse, ta disponibilité à discuter même quand ton agenda était plus que chargé, pour ton écoute et ta compréhension, pour trouver toujours une solution et pour me soutenir dans toutes les activités de vulgarisation.

Je voudrais aussi remercier Angelo, sans qui cette thèse n'existerait même pas. J'ai eu la chance de t'avoir comme co-encadrant pendant la première année. Ta présence et ton accompagnement pendant les premières phases de la thèse m'ont permis d'acquérir les compétences nécessaires pour continuer ce chemin. Merci pour tes conseils, ton humour et pour ta disponibilité même après ton départ. Encore aujourd'hui, après chaque discussion avec toi tout est plus clair !

Un Merci très important pour moi va à Brian. Merci pour tes conseils utiles et intelligents que tu m'as toujours donnés, et à tout ton soutien que tu m'as offert pendant cette dernière année. J'ai la chance de profiter de ton expertise et de ta douceur. Sans ton consent form la thèse n'aura pas été la même...

Le travail présenté dans cette thèse est le résultat de plusieurs collaborations et d'un travail d'équipe. Je voudrais donc remercier tous les collègues et les chercheurs qui ont contribué à ce projet.

Paul, tu as été mon premier stagiaire et tu seras toujours le "stagiaire" de référence (mes futurs stagiaires auront du mal à être à la hauteur :P). Merci pour tout ton engagement, surtout face aux problèmes techniques, et pour avoir partagé ensemble des semaines des passations qui des fois semblaient infinies !

Merci à Donatella, tu es arrivée au bon moment et ton aide a permis de bien conclure les expériences à la Cité des sciences. Merci pour ta bonne humeur, ta force et ton encouragement. J'ai passé des super moments avec toi. Malheureusement tu es partie sinon tu aurais fini par me convaincre de faire du sport :P

Merci à Soumia, avec qui j'ai travaillé sur les dernières parties de l'expérience. Merci d'avoir partagé les joies et les peines des démos et de l'administration ainsi que pour les discussions et tes conseils.

Je voudrais remercier l'équipe de Genève, notre partenaire du projet : merci à Chen pour ton travail et ton aide, j'ai passé des bons moments à Genève et à Paris ; merci à Guillaume et Thierry pour vos conseils et votre support.

Merci à Maurizio pour tout le soutien technique et pour la rédaction des articles.

Je voudrais aussi remercier Giovanna et Salvatore pour le soutien à mes travaux, les discussions, les conseils et pour me faire sentir en Italie.

Il ne faut pas oublier l'importance de l'environnement de travail : j'ai eu la chance d'avoir des collègues super cool tout au long de mon expérience dans la Greta team. Chacun de vous m'a apporté beaucoup, au niveau professionnel et personnel. Merci à vous tous : Florian, Nadine, Thomas, Guillaume, Yuko, Irina, Caroline, Valentin, Chloé, Léo, Soumia, Reshma, Sooraj, Fajrian, Tanvi, Nawal, Julien, Fabien.

Merci en particulier à Brice pour ta gentillesse, pour m'avoir aidé avec les soucis informatiques et pour toujours m'expliquer les mystères de VIB.

Et merci à Asma pour avoir relu et corrigé mon résumé en français :)

Je voudrais aussi remercier les secrétaires de l'ISIR, Anne-Claire, Sylvie, Michèle et Awatef, qui ont été toujours très gentilles et efficaces.

Merci à l'équipe de médiation scientifique de la Cité des sciences et de l'industrie, en particulier à Aurélie, Nadège et Olivier, qui m'ont accueillie lors de ma mission doctorale. Ça a été un plaisir de collaborer avec vous et découvrir le monde de la médiation scientifique.

Merci aussi à l'équipe du Carrefour Numérique, en particulier Laurence, Camille et Hélène, qui m'ont hébergée pendant les expériences. J'ai trouvé beaucoup de disponibilité et de sympathie.

Merci aussi à Florence Muri, François Xavier Jollois et Magali Champion de l'IUT Paris Descartes qui m'ont accueillie dans leur équipe d'enseignement. Merci de m'avoir fait confiance et de vos conseils pour gérer des étudiants des fois un peu difficiles, et de m'avoir donné la liberté d'organiser mes enseignements au mieux.

Je voudrais également souligner que je n'ai pas passé 3 ans et demi enfermée au labo mais j'ai pu aussi partager de bons moments avec mes amis en France et en Italie. Je voudrais remercier ici les personnes qui ont bien aimé discuter de ma thèse de temps en temps et qui m'ont supportée dans ce parcours.

Merci à mes collègues de l'orchestre OSIUP : Marc, Shadai, Julie et Philippe. Les discussions pendant les pauses et les trajets en bus ont été très sympathiques.

Merci aux membres du groupe “La crisi esistenziale" : Lisa, Federica, Anna, les 2 Francesche, Eugenia, pour tous les moments de partages de nos "crises" et de quelques moments de succès.

Merci à la Francesca Torino : merci pour m'avoir accueillie sans préavis quand j'ai eu besoin, pour ta gentillesse et ta douceur.

Merci à Paola pour tous les Skypes et nos tours de Trento au froid et au chaud.

Merci aux Luzzaresi, pour m'accueillir toujours chaleureusement et pour partager les plaisirs de la cuisine italienne à chaque fois que je rentre.

Pour finir, last but not least, un grand merci à mes parents. Merci pour tous les aller-retour à l'aéroport ou à la gare, avec des valises plus grandes que moi. Merci aussi pour vos visites dans la ville lumière. Cela représente seulement une petite partie de tout le soutien que vous m'avez donné pendant ses années.

Contents

List of Tables

List of Figures

Part I

Introduction

Chapter 1

Context of the Thesis

> Setting goals is the first step in turning the invisible into the visible.
>
> *Tony Robbins*

Contents

THIS Chapter introduces the context where this Thesis is placed, in particular the research questions we addressed and the approach followed to answer to these questions. The field of Embodied Conversational Agents is briefly introduced, as well as the motivation of this work and the main steps we realised to accomplish it. A list of the contributions of this Thesis and the publications who resulted from it is also given. The Chapter finished with the summary of the different Parts of the Thesis.

1.1 Introduction

Interaction is a fundamental need of human beings. In everyday life there are many occasions to interact with different people ranging from strangers to very intimates, such as a partner or a member of the family. When we meet strangers, the first moments are critical,

since we often form impressions about others that can have important consequences such as commitment to meet in further encounters, success at job interviews or dating again a potential partner (Ambady and Skowronski, 2008).

Goffman et al. (1978) define *impression formation* as the process of information perception, organization and integration in order to form coherent impressions of others (e.g., in terms of personality and interpersonal attitudes). We, as people, are aware of these mechanisms and we often attempt to control the impression that others form of us. This latter process is called *impression management* (Goffman et al., 1978), which mainly concerns the control of visual appearance (e.g. hair style, clothing). However, we also attempt to control social behaviour, but it may be difficult to have full control over all the social cues that are exhibited during the interaction. In particular, non-verbal behaviours are crucial because they can reveal with high accuracy a variety of information about us including, for instance, sexual orientation (Ambady and Skowronski, 2008), personality and interpersonal attitudes (Rosenberg et al., 1968).

During the last decades, anthropomorphic interfaces, such as humanoid robots and virtual characters, have been increasingly deployed in several roles, such as pedagogical assistants, companions, trainers. When conceiving Embodied Conversational Agents (ECAs), which are anthropomorphic virtual characters capable of interacting with users using verbal and non-verbal behaviour such as gestures, facial expressions and speech (for more details, see Cassell (2000)), it is very important to take into account how users perceive them during the course of the interaction. Virtual agents ought to be endowed with the capability of maintaining engaging interactions with users (Sidner and Dzikovska, 2005). This would make it easier for a virtual guide to transmit info, would ensure change behaviour for a virtual coach, would create rapport with a virtual companion.

Like in human-human interactions, the first moments of an interaction with ECAs are critical since users form impressions about them, that can affect the rest of the interaction, in terms of engagement and willingness to continue it (Cafaro et al., 2016). By managing non-verbal behaviours exhibited by a virtual agent we may improve the first impression about it given to the user. In this Thesis, we use the term "agent's impression" to refer to the impression given by the agent in order to be perceived by the user with different levels of warmth and competence (we introduce these variables in the next paragraph). We use the term "user's impression" to refer to this user's mental representation of agent's warmth and competence.

The goal of this Thesis was to build an ECA able to make the best possible first impression on a user, thus effectively engaging him or her in an interaction. This goal was realized by building an interactive loop which tied the behaviour of the agent to the actual reaction of the user facing it in real-time. We focused on the identification and modeling of the nonverbal behaviour, towards exhibiting, managing and maintaining impressions of two important socio-cognitive dimensions in the first minutes of interaction with a user, i.e. warmth and competence (W&C, (Fiske et al., 2007)). These dimensions are described

in Chapter 3. A schematic representation of the interactive loop is shown in Figure 1.1. In this schema, the agent plans the non-verbal multi-modal behaviours (such as smile, gaze, gestures, etc.) to display. These behaviours elicit in the user some impressions of the agent, in terms of warmth and competence. These impressions can be detected by the agent through the analysis of user's multi-modal behaviour. According to these impressions, the agent adapts the next behaviour to exhibit.

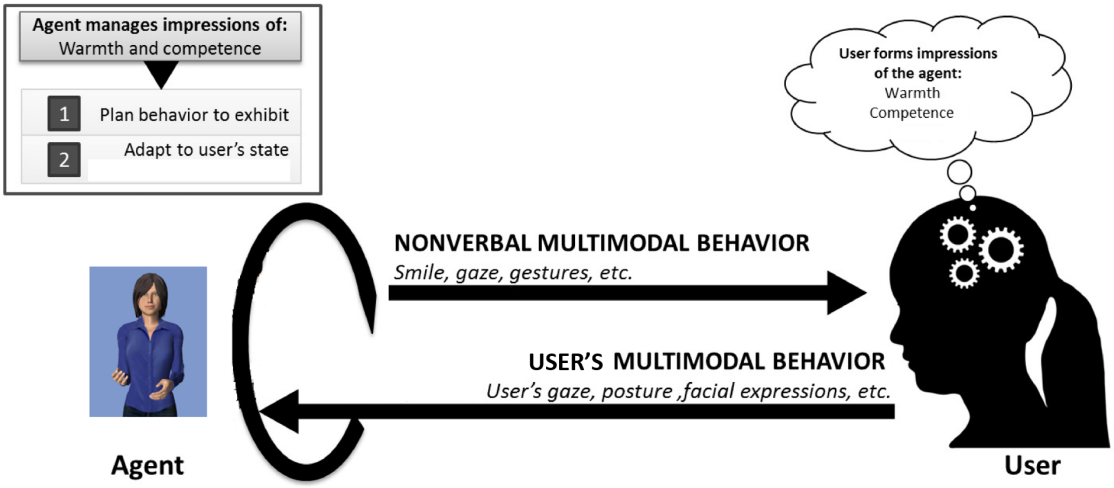


Figure 1.1 – A schematic representation of the computational model for impressions management developed during this Thesis. The agent manages impressions of W&C on the user by exhibiting nonverbal multi-modal behaviour and adapting this behaviour to the detected user's behaviour during the interaction.

The work presented in this Thesis has been realised in the context of the ANR project IMPRESSIONS, in collaboration with the Multimodal Interaction Group of the University of Geneva. It included external collaboration with Professor Maurizio Mancini and internal collaboration with Paul Lerner and Soumia Dermouche.

1.2 Embodied Conversational Agents

In artificial intelligence, an intelligent agent (IA) is an autonomous entity which observes through sensors and acts upon an environment using actuators (i.e. it is an agent) and directs its activity towards achieving goals (i.e. it is "rational", intelligent).

Intelligent software agents may also learn or use knowledge to achieve their goals. They may be very simple or very complex: a reflex machine such as a thermostat is an intelligent agent, as is a community of agents working together towards a goal (Russell and Norvig, 2016).

Different types of intelligent agents exist, according to their level of autonomy, interaction with the user, and other characteristics. Embodied Conversational Agents are those towards which research is increasingly focusing, and they are used in several domains. In this Thesis we will use the terms “ECA", “agent”, “virtual character” interchangeably to refer to ECAs.

An Embodied Conversational Agent is a computer-generated animated character that is able to carry on natural human-like communication with users (Cassell, 2000; Urbain et al., 2009). According to their name, ECAs are:

- *Embodied*, thus a personification of the machine, in the form of a physical body situated in the virtual environment, or even only an animated talking head. This characteristic distinguishes them from chatbots, which only perform verbal behaviour;
- *Conversational*, thus interactive, social, capable of engaging in conversations with one another or with humans employing the same multi-modal interaction means (Natural Language Processing, non-verbal behaviours, etc...), that humans do;
- *Agent*, since they maintain the characteristics of virtual agents, such as rationality, proactivity, autonomy. This characteristics distinguish them from avatars, which are a personification of a real human who controls them (Pauchet and Sabouret, 2012).

ECAs can be developed in different environments, including smart objects and virtual reality; they are modelled from human data gathered from both annotated corpora and related works like psychological knowledge. Figure 1.2 shows some examples of ECAs with different levels of sophistication and different roles.

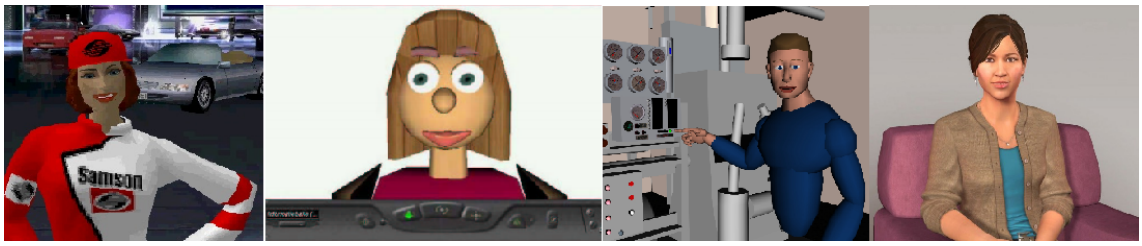


Figure 1.2 – Examples of ECAs. From left to right we see: Jennifer James, a car saleswoman designed by Extempo Systems Inc. who attempts to build relationships of affection, trust and loyalty with her customers (Elliott and Brzezinski, 1998); Karin, informing about theatre performances and selling tickets (Heylen et al., 2001); Steve, educating a student about maintaining complex machinery (Johnson and Rickel, 1997) and Simsensei, a virtual human interviewer for healthcare decision support (DeVault et al., 2014).

The Computer Are Social Acts (CASA) paradigm (Nijholt, 2002; Reeves and Nass, 1996), states that humans respond to computers as if they were social entities, thus at-

tributing attitudes and personality traits to them, as well as behaving following social conventions and reacting to them as they would react to a human. According to this paradigm, ECAs would be a more powerful way for humans to interact with their computers since they use a primary and early-learned skill of humans, i.e. conversation, that is a very defined characteristic of humanness and humans' interactions (Cassell et al., 2001). Face-to-face conversation applied to human-computer interfaces is often taken into consideration by designers, indeed this type of interaction allows for much richer communication, and enables pragmatic acts such as turn taking. Moreover, users have been found to prefer non-verbal visual indication of an embodied system's internal state to a verbal indication (Marsi and van Rooden, 2007). Endowing ECAs with the ability of exhibiting the appropriate non-verbal behaviours during the interaction with the user has been the goal of many researchers in the last decades. They mainly focused on the ECA's expression of emotional states (Pelachaud, 2009), personality traits (McRorie et al., 2009) and interpersonal attitudes (Ravenet et al., 2013b) via non-verbal behaviour.

1.2.1 SAIBA Framework for ECAs

In order to stimulate human-like multi-modal communicative behaviour, advanced ECA systems need to incorporate a whole range of complex processing steps, from intent to behaviour planning to behaviour realization including some sort of scene or story generation, multi-modal natural language generation, speech, synthesis, the temporal alignment of verbal and non-verbal behaviours, and behaviour realization employing particular animation libraries and engines.

In this Thesis we exploited the Greta/VIB Platform (Pecune et al., 2014) for the real-time generation and animation of ECA's verbal and nonverbal behaviours. The platform follows the SAIBA framework which organises the different processing steps for behaviour generation into three top level modules, as shown in Figure 1.3 (Krenn et al., 2011; Kopp et al., 2006; Vilhjálmsson et al., 2007):

1. *Intent Planner*: this high-level module concerns agent's communicative intentions, i.e. its goals, emotional states and beliefs, that affect what the agent wants to communicate to the user. These intentions are encoded with Functional Markup Language (FML). It can also include a dialogue planner, which generates an initial abstract version of a dialogue, as a sequence of dialogue acts, without specifying the words, but only the communicative functions of the dialogue, such as requesting for information, answering a question, giving feedback, etc... Thus, at this level no reference to any physical or verbal behaviour is specified.

2. *Behaviour Planner*: this low-level module conveys the output generated by the previous module (i.e. the agent's intentions), by scheduling a number of communicative signals, such as speech, facial expressions and gestures, encoded with Behaviour

Markup Language (BML), an almost standardized language commonly used by ECAs community. BML language specifies the verbal and non-verbal behaviours, independently of the particular realization (animation) method used. In addition, information about the temporal order and the relative dependencies between the communication channels involved is generated.

In other words, the Behaviour Planner gives a detailed description of the behaviours (verbal and non-verbal) that should be performed by the agent in order to convey its intentions. These descriptions are written using a specific language, BML, which contains tags specifically aimed to describe behaviours.

3. *Behaviour Realizer*: the goal of this last module is to realize the behaviours scheduled in the previous planner by generating animation from the BML inputs. This task involves a series of steps like the synchronization of different modalities and the resolution of conflicts between behaviours of the same modality.

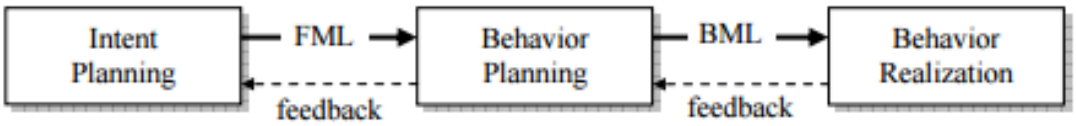
   The main important outputs of Behaviour Realizer are:

   - Speech generation: thanks to an additional module, sounds files are generated given input text annotated with additional information. Pronunciation, prosodic properties, pitch, duration of phonemes can be generated, as well as a temporal information to allow for multimodal synchronization, for instance with visemes (mouth shapes related to sounds/phonemes) and with gestures;
   - Gestures realization: gestures are composed starting from basic hand-arm movement trajectories, defined according to the position of the wrist in 3D space, and the hand configuration. More complex gestures can be built by sequences of basic gestures;
   - Facial expressions: there exist several coding schemes to describe facial expressions. For example, Facial Animation Parameters (FAPs) are used to represent basic facial actions including tongue, eye and mouth movements, that can combine to represent facial expressions.

1.2.2 Applications of ECAs

ECAs are suitable for several different contexts and roles (Pauchet and Sabouret, 2012):

- Assistant: this type of agent has the function of welcoming and assisting the user in understanding and using an application or a website. It has the capacity to resolve help requests (often inputs in natural language or speech) issuing from users with poor knowledge about a component (application, website, etc.). Assistant agents are often present in e-commerce, or in websites which offer assistance to customers.



.

Figure 1.3 – The three stages of behaviour generation in the SAIBA framework and the two mediating languages FML and BML (Vilhjálmsson et al., 2007)

- Tutor: well-built ECAs can be useful for students in the context of e-learning, but also for patients, in monitoring systems, both for psychological and physiological problems.

- Partner: this role is thought for actors in virtual environments such as games, virtual communities, working groups. The ECA in this context can help for example to solve a problem cooperatively or to moderate a virtual meeting or a teleconference.

- Companion: ECAs can be used in some cases without specific functions, but only to become a virtual friend in long term relations.

The beginning of the interaction is a typical phase for all of the applications listed above. For instance a virtual teacher would need to quickly ensure a good impression of competence in front of learners, while a virtual barman would need to display warmth in the short duration of a drink order. All these systems could benefit from a good management of the first moments of the interaction. There has been extensive research on how to improve the user-agent level of engagement during the interaction. However, first impression management has been neglected. Except for a few studies focusing on first encounters (Cafaro et al., 2012, 2016) and the impact of nonverbal realization choices on users' impressions (Ter Maat et al., 2010), little is known about the importance and the benefit of managing first impressions on users in the initial formative phases of the human-agent interaction. For example, the Avatar 1:1[1] project focused on maintaining user's engagement with a museum agent, but it assumed that the first contact with the user had been already established. In the FP7-NoE SSPNet[2] project only user's social signals where considered, without taking into account their possible relation with first impressions.

That's why in this Thesis we focused on improving ECAs capabilities to manage their impressions during the beginning of the interaction taking into account user's reactions.

[1] http://lifesizeavatar.com/
[2] http://sspnet.eu/

1.3 Research Questions and Approach

The motivation of the work presented in this Thesis was to improve the quality of agent's impressions generated on user and user's engagement in human-agent interaction. In particular, we focused on the virtual agent and addressed the following research questions:

- (**RQ1**) *How to **model** non-verbal behaviours linked to warmth and competence in a virtual agent?*
- (**RQ2**) *How to **adapt** the virtual agent behaviours to the impressions formed by users?*

The work presented in this Thesis was placed in the context of ANR IMPRESSIONS Project, in collaboration with the Multimodal Interaction Group of the University of Geneva. The project addressed also the following research question: "*How to **measure** impressions from users' behavioural expressions?*". This was the topic of the PhD Thesis of Chen Wang. We collaborated with her and integrated her model for detecting user's impressions about the agent in a real-time interaction. This work is reported in Chapter 9.

To address **RQ1** and **RQ2**, our approach started from investigating whether the same processes that characterize human social cognition also apply in human-agent interaction. We started from the analysis of a corpus of mediated natural human-human interaction, aiming at finding non-verbal cues eliciting different degrees of W&C, with the purpose of applying these findings in human-agent interaction. Our approach was centered on non-verbal behaviour, since, as we mentioned in the previous Section, this modality is very important in impression formation processing and it is possible to control and to manipulate it in the virtual agent.

Our approach followed 4 main steps:

Step 1: W&C impressions about humans

The first step consisted in analyzing a corpus of dyadic expert-novice knowledge sharing interactions, with the purpose of building a repertoire of non-verbal signals eliciting different degrees of W&C impressions. This step aimed at answering these specific questions:

- ($\mathbf{q1_a}$) *Can non-verbal behaviour affect people's impressions of W&C?*
- ($\mathbf{q1_b}$) *If so, what are the non-verbal cues associated to these impressions?*

This work is described into details in Chapter 5.

Step 2: W&C impressions about virtual agents

As a second step, the findings of the previous one were implemented in an ECA and manipulated in order to investigate whether they were perceived in the same way when expressed by an ECA. This step aimed at answering these specific questions:

- (**q2$_a$**) *Is a virtual agent perceived differently in terms of W&C according to the non-verbal behaviours it realises?*

- (**q2$_b$**) *If so, what are the non-verbal cues (or combinations of non-verbal cues) that allow it to be better perceived in terms of W&C?*

- (**q2$_c$**) *Do our expectations and a-priori of an ECA influence the impressions that are formed afterwards?*

This work is described into details in Chapter 6.

Step 3: System architecture for agent's impressions management

Based on the findings of the previous steps, the goal of this step was to answer to **RQ2** by building a system to manage agent's impressions in real-time during the interaction with a user. The architecture consisted in 3 main modules: (1) one for detecting user's reactions by processing low-level signals such as facial expressions, head rotation, etc.. into high-level variables such as the degree of user's impressions or user's engagement (see next paragraph); (2) one for agent's adaptation according to user's reactions; (3) one for agent's behaviour generation.

The aim of the adaptation module (2) was to endow the agent with the ability to cope with real-time user's reactions during the interaction, and to adapt its exhibited behaviour as a function of its own intents (i.e. managing impressions of W&C) and the user's formed impressions. The goal of behaviour adaptation was to yield to better impressions management and to be able to maintain the wanted impressions.

This architecture is described in Chapter 7.

Step 4: Use Cases

Once we implemented the general architecture for impressions management, we conceived two use cases where we applied the system to a real interaction scenario, in order to evaluate the impact of the impressions management model on user-agent interaction.

Engagement plays an important role in human-agent interaction: an engaging virtual agent is more likely to be accepted by the user, as well as to promote future interactions Bergmann et al. (2012), Cafaro et al. (2016). In the first use case, we focused on investigating how to adapt agent's behaviour according to user's engagement. The aim was to answer these questions:

- (**q4$_a$**) *Is there a relationship between W&C impressions and engagement during the interaction with an ECA?*

- (**q4$_b$**) *Is it possible to improve user's engagement by managing agent's degree of W&C?*

In the second use case, we exploited the impressions detection model developed by Chen Wang (Wang et al., tted) to adapt agent's behaviour according to user's impressions detected in real-time. The aim was to answer these questions:

- (**q4$_c$**) *Is it possible to influence user's impressions of agent's W&C by adapting agent's behaviour to user's impressions?*
- (**q4$_d$**) *Does agent's adaptation affect user's overall perception of the interaction?*

These works are described into details in Chapters 8 and 9.

1.4 Contributions

The scientific contributions of this Thesis can be summarized as follows:

First contribution: *Creation of a repertoire of multi-modal behaviours eliciting impressions of warmth and competence.*

The planning and realisation of the agent's multi-modal behaviour required a modeling phase of human inspired non-verbal behaviour typically exhibited for expressing the impressions of warmth and competence. Therefore, we aimed at defining a repertoire of human multi-modal behaviour that could be modeled into the agent to manage the first impressions of the selected dimensions. This repertoire was built by starting from the study of literature about non-verbal cues of W&C (Cuddy et al., 2011; Bayes, 1972; Maricchiolo et al., 2009). We integrated the existing findings with the analysis of a corpus of natural human-human interactions and we investigated the effect of these behaviours in virtual agents perception.

Second contribution: *Creation of a behaviour adaptation module for ECA's impressions management.*

We created a module based on reinforcement learning allowing the ECA to adapt its behaviour to user's reactions to enhance an interaction loop. In particular, the set of possible behaviours that the agent could choose to perform derived from the previous steps about the analysis of human-human interaction and agent's perception. The agent could learn to adapt its behaviours in real-time without having previous knowledge about user's reactions to its behaviour. It would start by exploring the effects of its behaviour and then, once obtaining enough knowledge about it, it could select the behaviour best fitting its goals.

Third contribution: *Creation of a set of strategies for managing impressions of warmth*

and competence in an ECA.

Starting from the findings coming from the analysis of human-human interaction, we investigated the role of multi-modal behaviours and expectancies when judging virtual agents, in order to create a set of strategies for managing impressions of warmth and competence in an ECA. These strategies were implemented in the general architecture for agent's impressions management, allowing it to choose how to manipulate its behaviour during the interaction.

Fourth contribution: *Implementation of the impressions management module in a system architecture for real-time user-agent interaction.*

We integrated the impressions management module in a human-agent system including a module to detect and interpret user's multi-modal reactions and a module for agent's behaviour generation. The architecture is general enough to allow for customisation of the different modules according to different contexts and goals of the agent.

Fifth contribution: *Investigation of the effectiveness of an adaptive agent and of the relationship between agent's adaptation, engagement and impressions, during human-agent interaction.*

We conceived a scenario where the virtual agent played the role of virtual museum guide and we personalised the general architecture for impressions management in 2 different applications. In the first one the agent adapted its self-presentational strategies to be perceived more or less warm or competent with the goal to maximise user's engagement. In the second case it adapted its behaviour in order to maximise user's impressions of its warmth or competence level. We designed and conducted an evaluation study for each scenario in order to validate the effectiveness of the agent's impressions management capabilities. In particular we wanted to verify that users preferred an agent with impressions management skills over an agent which did not manage users' impressions.

1.5 Publications and Dissemination

This Thesis gave rise to several national and international publications, as well as to invited talks and dissemination to non-expert audience. Additional discussions about my PhD research occurred in the context of the *ISSAS 2018* Summer School[3] and several edi-

[3] https://www.unige.ch/cisa/education/summer-school-issas-2018/

tions of the *SMART School on Computational Social and Behavioural Sciences*[4].

The list of publications and dissemination can be found in Annex A.

1.6 Thesis Structure

This Thesis is organised into 7 parts, including this Introduction. In Part II we present the theoretical background about first impressions (in Chapter 2) and warmth and competence dimensions (in Chapter 3).

In Part III we discuss related work about ECAs, in particular the studies which included W&C dimensions in a virtual agent, studies conducted in the Museum fields and studies involving engagement (in Chapter 4).

Part IV is devoted to the first step of our approach, which consisted in the analysis of W&C impressions in human-human interaction. In Chapter 5 we introduce the corpus analysis and the findings about non-verbal behaviours associated to different W&C levels.

In Part V we focus on the perception of W&C in virtual gents, by describing the perceptual study we conducted in order to investigate the effect of non-verbal behaviours on W&C impressions when displayed by an ECA instead of a human (Chapter 6).

Part VI is devoted to the study of W&C in human-agent interaction. Our architecture for agent's impressions management is described in Chapter 7. The applications of this architecture to different use cases and their evaluation studies are presented in Chapters 8 and 9.

Finally, Part VII resumes the contributions of our work as well as its limits, and propose some perspectives to improve it in a short and long term (in Chapter 10).

[4] http://www.smart-labex.fr/SMART_School_on_Computational_Social_and_Behavioral_Sciences.html

The key points of this Chapter:

Research Questions:

- *How to **model** non-verbal behaviours linked to W&C in a virtual agent?*
- *How to **adapt** the virtual agent behaviours to the impressions formed by users?*

Goal of this Thesis:

- to build an ECA able to make the best possible first impression on a user, thus effectively engaging him or her in an interaction.

How to realize it:

- by starting from the analysis of human-human interaction;
- by building a reinforcement learning loop to adapt the behaviour of the agent to the actual reactions of the user interacting in real-time.

Part II

Theoretical Background

Chapter 2

First Impressions

> A thousand words leave not the same deep impression as does a single deed.
>
> *Henrik Ibsen*

Contents

FIRST impressions are a psychological process studied from the 40's (Asch, 1946) and then along the 60's (Anderson, 1962). More recent models of impression formation have been developed, among them one of the most prominent ones is

the continuum model by Fiske and Neuberg (1990). A common aspect of those models is that forming an impression about a stranger can be described in three steps. First, we *perceive* the person newly met, often visually, as vision is the fastest sensory channel. People immediately perceive and collect information about invariants traits such as age, sex, ethnicity, and variant traits, such as face, gestures, body posture, gaze. After acquiring this first information, people make *inference* about the other, for example about his/her personality. Finally, the new person is *categorized* in a certain group or subgroup in which the perceiver either feels to fit in or not. Stereotypes often influence impression formation processes (Fiske et al., 2002). In this Chapter we review the main psychological theories about the two main components of first impressions, as labeled by Goffman et al. (1978): impression formation and impression management. Some of the proposed frameworks can be applied to our research.

2.1 Introduction

When we meet strangers, the first moments are critical, since we often form impressions about others that can have important consequences such as commitment to meet in further encounters, success at job interviews or dating again a potential partner (Ambady and Skowronski, 2008).

When talking about first impressions, Goffman's notion of impression management and formation (Goffman et al., 1978) is central:

- *Impression formation*. It is the process by which individuals perceive, organize, and ultimately integrate information to form unified and coherent situated impressions of others. Internalized expectations for situated events condition what information individuals deem is important and worthy of their attention. Further, these expectations condition how individuals interpret this information and serve as the basis for subsequent attributions (Moore, 2007).

- *Impression management*. Also called self-presentation, impression management refers to the process by which individuals attempt to control the impressions that others form of them. People, whether or not thinking about it, are often engaged in impression management, trying to control the information that others receive about them (Miller, 2010; Goffman et al., 1978).

In this Chapter we provide theoretical background about these two areas and identify some elements of these theories that we took into account in the research work presented in this Thesis. The Chapter is organised as follows: in the next Section, we describe the general steps involved in the process of impression formation; in Section 2.3 we present the main approaches that attempted to explain how impression formation works; in Section 2.4 we describe an interesting model about the factors that affect impression management.

2.2 Impression Formation

The study of Impression Formation started to develop in connection with the progress in the measurement of subjective phenomena, such as personality and attitudes. Quantitative methods like the Likert scale, paired comparisons and factor analysis made the study of subjective phenomena more "scientific", so social psychology became more experimental and started to consider subjective concepts such as personality traits as "objects of perception".

The first who investigates how people form impressions about others was Asch (1946), who applied a simple paradigm where he presented to participants a list of traits to describe a hypothetical person and asked them to give their impression about this person. After him, many other researchers addressed to this research area by proposing different theories about impression formation. We will present the most relevant ones in Section 2.3.

In Moore (2007) impression formation is defined as "the process by which individuals perceive, organise, and ultimately integrate information to form unified and coherent situated impressions of others". This definition highlights 3 steps which lead to the final impression of others: information processing, attribution and information integration. We describe them into details in the next paragraphs.

2.2.1 Information Processing

In order to judge others, we need information about them. The source of this information could be the target person directly or an indirect information given by another person.

In human-human interaction first impressions can be formed by observing individual characteristics such as height, clothing and, more generally, visual appearance (Naumann et al., 2009; Argyle, 1975; Miller, 2010). However, impressions can be formed also by observing someone behaviour, such as facial expressions and body language (i.e. nonverbal behaviour) (Riggio and Friedman, 1986; Argyle, 1975; Burgoon et al., 1984; DePaulo, 1992).

Information processing is affected by our cognitive limitation. Indeed, we cannot pay attention to all the available information, thus we have to select the one which is worth to be gathered. Internalized expectations coming from previous experience condition this process. An example is the *confirmation bias* (Plous, 1993), that occurs when "individuals tend to search for, interpret, favor, and recall information in a way that confirms their preexisting beliefs or hypotheses".

2.2.2 Attribution

After having collected information about the other, people tend to find the causes of these behaviours, in order to better understand them and to predict future behaviours.

Attribution theories (e.g., Heider (2013); Jones and Davis (1965); Kelley (1967)) suggested that people understand each other's behaviour as arising from two causes: dispositions and situations. Dispositional attributions, also called internal attributions, refer to the process of assigning the cause of a behaviour to some internal characteristics, like ability and motivation, rather than to outside forces. Situational attributions, also called external attributions, refer to interpreting someone's behaviour as being caused by the context where the individual is.

From this perspective, predictions about others' behaviour could be accomplished via simple algebra: disposition + situation = behaviour (Lieberman et al., 2002). In the real world the relationship between situation and disposition is not always straightforward so people could fall into attribution biases. Some of them are: the fundamental attribution errors (Jones and Harris, 1967) when we tend to rely only on others' dispositions; cultural bias (E. Dent, 1974) when attributions rely only on others' own culture; self-serving bias (Miller and Ross, 1975) when we attribute success to dispositions and failure to external factors.

2.2.3 Information Integration

The final step of impression formation consists in combining all the information and attributions collected about the target in order to produce a final impression.

Different theories attempted to give explanations about how information integration works. We describe the most important approaches in the next Section.

2.3 Theories about Impression Formation

Starting from 40's, several social cognition theories provided explanations of the process of general information gathering and processing. Two major approaches emerged form this line of research. One approach considered impressions in their globality (by following the principles of Gestalt psychology), as the result of the relation between the individual traits. This was firstly proposed by Asch and is described in Section 2.3.1. The second approach applied mathematics rules to compute the final impression from the values of individual traits. It was proposed by Anderson and it is described in Section 2.3.2. More recently, other approaches emerged, which integrated cognitive and motivational elements in the process of impression formation. The most relevant one was the model of Fiske and Neuberg (1990) which highlighted the role of stereotypes, personal motivation and attention resources in the formation of impressions. Their theory is described in Section 2.3.3.

2.3.1 Asch's Gestalt Model

Asch (1946) was the first who initiated the study of impression formation. He focused on how people form impressions by judging sets of personality-traits used to describe a hypothetical person. In Asch (1946) he investigated different theoretical frameworks to explain impression formation. The first one suggested that we form impressions about single traits, and the total impression of a person is the sum of these independent impressions. In a variant of this framework (showed in Figure 2.1), another factor, called general impression, defined as "an affective force possessing a plus or minus direction which shifts the evaluation of the several traits in this direction" was added to the initial sum (this factor is related to the *halo effect*, see Section 3.4).

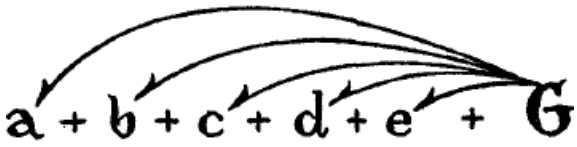


Figure 2.1 – The representation of an impressions according to the first framework investigated by Asch (1946).

The second approach, inspired by Gestalt psychology, suggested that we form a unified impression of the person that is not the sum of each trait, but rather the perception of a particular form of relation between these traits, as depicted in Figure 2.2.

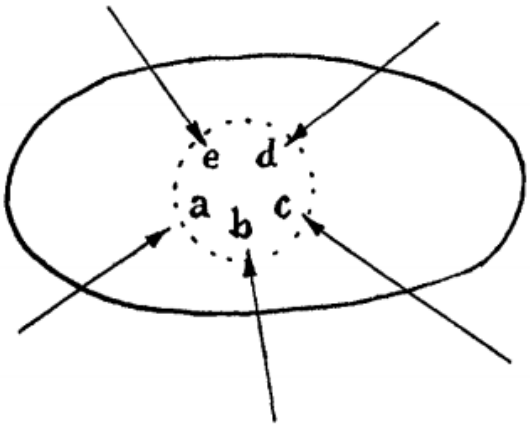


Figure 2.2 – The representation of an impression according to the second framework investigated by Asch (1946).

To check which of these frameworks was the most appropriate to describe the process of impression formation, Asch conducted a series of experiments where participants judged a hypothetical person given a list of traits (this study is described into details in Section 3.2). The results supported the Gestalt model: forming an impression was found to be an organized process where traits were perceived in their dynamic relations. That is,

isolated traits lost their individuality and entered into a structure where they had a particular relation with each other. Thus a change in a single trait would modify the entire impression. In addition, Asch found that traits did not have equal weight: he distinguished among central and peripheral traits. Central traits were those which had a strong effect on the perception of the other traits, while peripheral traits did not had the same influence on the global impression.

2.3.2 Anderson Algebraic Model

In contrast to the Gestalt approach of Asch, Anderson (1962) proposed an algebraic approach to explain impression formation processes. According to his model, we assign numeric values (positive and negative) to each trait that we encounter in a person. Thus, individual traits are evaluated independently, and the final impression of the target person consists in the weighted sum of these values.

According to his model, the final impression was defined by the following formula:

$$I = \frac{w_{\mathrm{o}} s_{\mathrm{o}} + \sum_{k=1}^{N} w_{\mathrm{i}} s_{\mathrm{i}}}{\sum_{k=1}^{N} w_{\mathrm{o}} + s_{\mathrm{o}}} \tag{2.1}$$

where:

- N is the number of single traits attributed to the person;
- s_{i} is the scale value, i.e, the numeric value assigned to the trait i;
- w_{i} is the weight, i.e., the functional importance of the trait i in the rating process;
- w_{o} and s_{o} represent the scale value and the weight of the initial impression, prior to receiving any information about the target.

In a series of studies (e.g., Anderson (1968); Anderson and Alexander (1971); Hendrick (1968); Lampel and Anderson (1968); Oden and Anderson (1971) the proposed formula was shown to account for very extensive sets of social judgment data.

Both Gestalt and algebraic models presents some weaknesses. Indeed, they did not focus on the role of contextual information and did not explain the processes involved in impression formation. In the next subsection we present a more recent model which focused more on these variables in the dynamic process underlying impression formation.

2.3.3 Continuum Model

According to this model (Fiske and Neuberg, 1990) impression formation is a dynamic process. People develop impressions of others by using a range of processes in a continuum

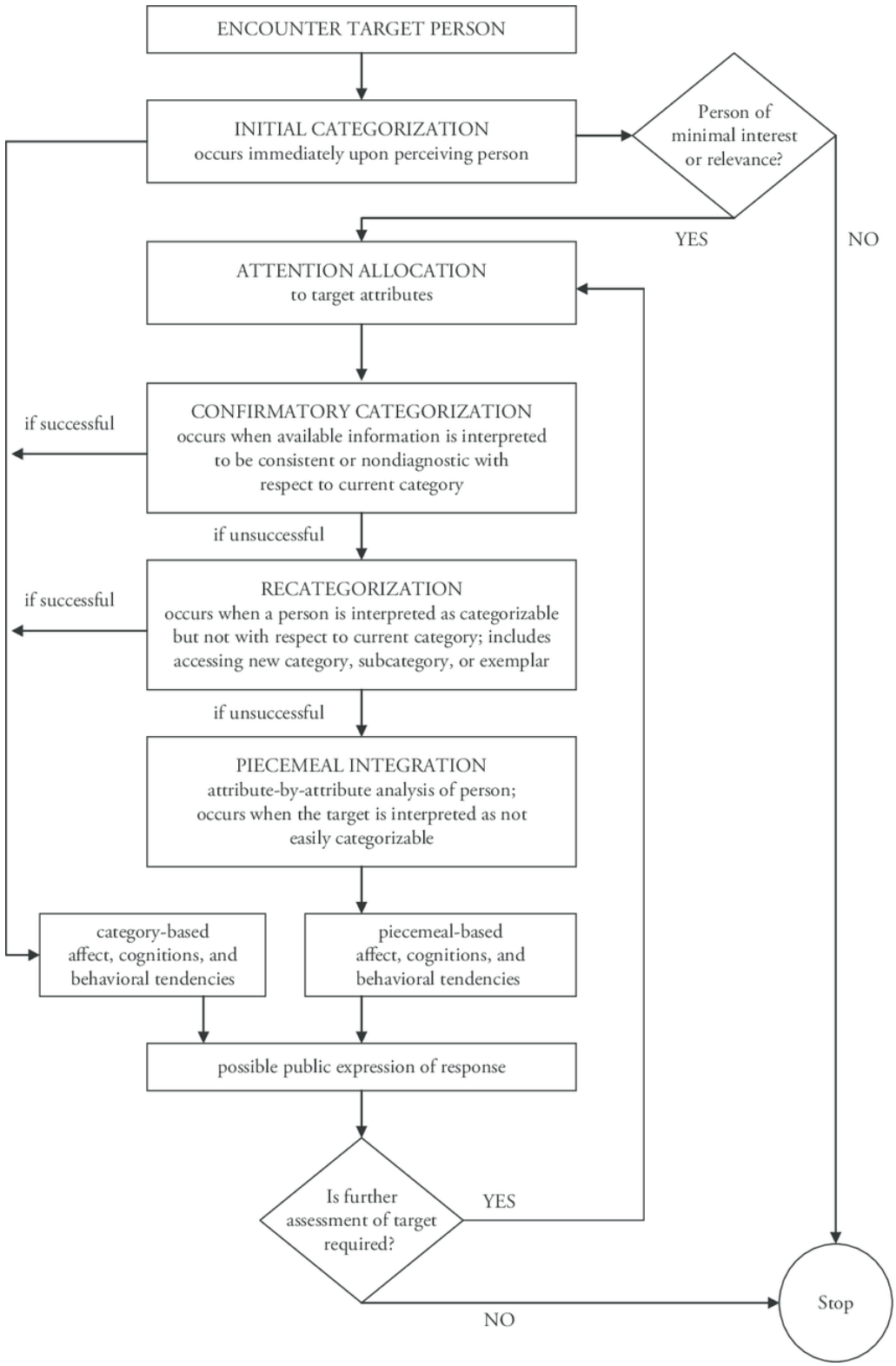


Figure 2.3 – The continuum model of impression formation (adapted from Fiske and Neuberg (1990)).

from fast categorization to piece-meal integration of target's attributes. The possible steps of the impression formation process, depicted in Figure 2.3, are:

- *Initial Categorization*: this is the default mode. As soon as we encounter a new person, we automatically place him/her into an existing social category. Influenced by stereotypes, we usually associate emotions, cognition and behaviour to this social category. According to several factors, we can decide to ignore stereotypes and go further into the continuum to adjust our impression.

- *Personal relevance*: this is one of the factors that determine the motivation to continue to gather information about the target, during each step of the continuum, in order to adjust the initial impression. If at any time the target is no more relevant to us, no more information is collected and the impression stops to be adjusted. For example, we would be motivated to go further in the continuum for targets with whom we have to accomplish a share goal, while we would prefer to stop at the initial category if we want to justify power over the target.

- *Attention Allocation*: this is the second factor that determines whether or not to continue to collect additional information about the target. We must have the sufficient resources, such as the time and energy, to do this. If resources finish, no more information is collected and the impression stops to be adjusted.

- *Confirmatory Categorization*: it occurs when we try to assimilate additional information about the target that is consistent with the initial categorization. If this is successful, our attitude toward the target will be based on the initial categorization. If the target's traits are inconsistent with the initial category, we will continue with the next step of the model.

- *Recategorization*: it occurs when we try to find a new existing category for the target according to the additional information. If recategorization is successful, we form new attitudes based on the new category.

- *Piece-meal Integration*: if the target cannot be placed in any pre-existing category, in this phase we take into account all the attributes we collected about him/her and a new attitude is formed towards the target.

- *Public Expression*: the last step of the impression formation process occurs when we decide (either consciously or unconsciously) to express the formed attitude toward the target. This can occur at any time phase of the continuum. If it occurs at the initial steps, the expressed attitude can reflect prejudices and stereotypes.

In the work presented in this Thesis we took into account these theories since they could help us in the better conceiving our model for the agent. In our context it was

important to take into account people's a-priori that could influence the information processing step. Our model took inspiration from the framework of the continuum model of Fiske and Neuberg (1990), indeed users continuously collected information about the agent and uploaded their impression about it during the interaction. The agent did not know at the beginning how to behave and tested some behaviours that may not be the right ones. Thus our goal was to create a global impression into the users that took into account how the agent improved its behaviour during time, instead of a mere sum of all the positive and negative impressions formed during the interaction.

2.4 Impression Management

Besides forming impressions about others, people give high importance to "what the others think" about them thus they often try to control the impressions others form of them. As already defined at the beginning of this Chapter, Impression Management "describes efforts by an actor to create, maintain, protect, or otherwise alter an image held by a target audience" (Bozeman and Kacmar, 1997). Leary and Kowalski (1990) reviewed the main theories about the factors that influence impression management, in particular the social point of view of Goffman et al. (1978), the psychological approach of Jones and Pittman (1982) and other approaches such as that of Schlenker (1980). From this review they identified two main components underlying impression management and they proposed a model to explain their role in this process: *impression motivation* and *impression construction*.

2.4.1 Impression Motivation

The first factor that affects impression management concerns the desire to create one particular impressions in the others, but this does not necessary imply that this impression will be realized. According to the situation, people do not care about the impression they give to others, for example when they experience ecstatic joy (Duval and Wicklund, 1972). On the other hand, their level of self-awareness, that is, how people are conscious about being an object judged by others (Duval and Wicklund, 1972), can be extremely elevated, such as in case of a first date or a job interview, and increase the motivation to manage one's impression. People tend to consider other's impressions at a pre-attentive level unconscious level and to activate their impression management process at any time when they believe it is worthy.

In the two-component model of Leary and Kowalski (1990) three main factors are identified that affect impression motivation: the *goal-relevance of impressions*, the *value of desired goals* and the *discrepancy between the desired and the current image*. In particular, these factors contribute to the general goal to maximise expected rewards and

minimize expected punishments (Schlenker, 1980). This goal includes three distinct self-presentational motives:

- Social and material outcomes: giving the right impression would increase the possibility to obtained desired outcomes, both social, like friendship, power, etc. and material, like an increase in the salary;
- Self-esteem maintenance: others' reactions like compliments or criticism would improve or deflate our self-esteem;
- Development of identity: self-presentation is important to establish a desired identity, since identity is defined by society (Mead, 1934).

2.4.1.1 Goal-relevance of impressions

The motivation for impression management is affected by the level of relevance that conveying that impression has in achieving one's goal. This goal-relevance is related to how public the behaviour is: the higher is the number of people forming impressions about one's behaviour, the higher is the motivation to perform that behaviour (Baumgardner and Levy, 1987). It is also a function of the individual's dependency on the target: we are more motivated to give a good impression to our boss than to a friend (Bohra and Pandey, 1984). Finally, the importance to convey an impression to a target is also linked to the possibility of future encounters with that person (Schneider, 1969).

2.4.1.2 Value of desired goals

In addition to the relevance of an impression to achieve one's goal, our motivation to control our impression is affected by the relevance of the goal itself. The value of an outcome is related to how desirable it is, for example, we are more motivated to impression management if we apply for a good job than a job less interesting (Beck, 2003). The value of the outcome is also related to the characteristics of the target, for example our motivation increases for powerful or attractive targets (Schlenker, 1980).

2.4.1.3 Discrepancy between desired and current image

Impression motivation also increases when we realise that others' impression is distant from the desired one. For example, in case of failure or embarrassment, other people could have negative impressions about us, that do not correspond to our goal, so we are motivated to control our impression in order to limit this discrepancy between our desired impression and the current one (Frey, 1978; Brown, 1970).

In the context of our work, this factor represented the main motivation of our Embodied Conversational Agent: user's impressions were constantly monitored and when these

impressions did not match which the agent's goal, it would manage them by adapting its behaviour in order to elicit the desired impression into the the user.

2.4.2 Impression Construction

The second factor that affects impression management concerns the type of the impression that one aims to convey and how to realise it, that is, what means to use. We can control our impression in terms of personality traits, attitudes, roles, beliefs, etc. To do this, we can use self-description, non verbal-behaviour, physical appearance, association with social groups.

In the two-component model of Leary and Kowalski (1990) five main factors are identified that affect impression construction: they are described in the next paragraphs.

2.4.2.1 Self-concept

Impression construction is affected by how people think they are. People want to ensure that others accurately perceive their characteristics they are proud of. Through impression management they can tactically select specific characteristics to portray according to the particular target or situation. People hesitate to give impressions inconsistent with their self-concepts, that is, that differ too much from what they really are, both because of the risk to be "unmasked", and because of their internalised ethic against lying (Schlenker, 1980). People who tend to act self-presentation dissimulation and pretense, thus people whose self-presentation differs form their self-perception, are often those with high-variability roles, such as politicians, or involved in superficial relationships (Buss and Briggs, 1984).

2.4.2.2 Desired and undesired identity images

People have desires about the image they would like to be or not to be. This affects the impressions they convey, for example they can behave consistently to their desired identity and inconsistently to the undesired identity image (Schlenker, 1985).

2.4.2.3 Role constraints

Impression construction is also affected by the necessity to convey impressions consistent to one's role. For example, people in a position of authority have to maintain an impression of competence and leadership, in order to continue to be efficient in their role (Leary et al., 1986).

2.4.2.4 Target values

People can also select their characteristics that match with the preferences of the others. This strategy is not necessarily deceptive, for example people can try to omit information that does not fit with the target's values (Leary and Lamphere, 1988).

This factor was important in our context since the goal of our agent was to adapt its behaviour in order to match user's preferences. It had a set of possible behaviour to select, but the final choice took into account only those that were positively judged by the user.

2.4.2.5 Current or potential social image

The last factor that people take into account when creating their impressions is the image that others can have about them, currently and in the future. This makes people reluctant to convey impressions inconsistent with their self-perception, because they will have low probability to maintain this impression in the future. Another consequence of this factor is the compelling to use certain strategies such as face-saving or apologies to repair damaged social images, for example after public failures or embarrassing events (Snyder et al., 1983; Schlenker and Darby, 1981).

2.5 Non-verbal Behaviour in Impression Management

In the context of this Thesis, we were interested in the role of non-verbal behaviour in impression management. Relatively few works showed how people manage their non-verbal behaviour in order to control the impression to give to the others. For example, Rosenfeld (1966) conducted an experiment where participants interacted in dyads under 2 conditions. In the approval-seeking condition one of the two people in the dyad was asked to gain the approval of the other person, while in the approval-avoiding condition he was asked to avoid the approval of the other person. Participants in the approval-seeking conditions produced more smiles, head nods, gesticulations and verbal responsiveness than participants in the approval-avoiding condition. In addition, approval from the other member of the dyad was positively correlated with head nods, verbal responsiveness and negative correlated with self-manipulations and self-references.

The majority of the experimental studies about impression management revealed that self-presentation concerned the use of verbal behaviour rather than non-verbal behaviour. For example, in Peeters and Lievens (2006) participants who were instructed to convey a favorable impression used more proactive, assertive self-focused and other-focused verbal tactics than people who were instructed to convey an accurate impression. Impression management instructions had no influence on the use of non-verbal tactics. This suggests that non-verbal behaviour might be less intentionally controllable, probably because non-verbal reactions occur very fast and more spontaneously than verbal reactions. Indeed,

research evidence showed that attempts to produce specific non-verbal behaviours often cannot be executed successfully (DePaulo, 1992).

In our context, since we worked with Embodied Conversational Agents, we could control and manipulate their non-verbal behaviour. Our goal was to investigate how the management of agent's non-verbal behaviour can impact user's impression formation about the agent.

2.6 Conclusion

In this Chapter we presented the theoretical background about first impressions, in particular about the two main components of this phenomenon that are impression formation and impression management. Impression formation includes three steps: information processing, attribution of causes of others' behaviours and information integration. Different theories tried to explain how people form impressions about others. One approach considered impressions in their globality, as the result of the relation between individual traits of the other. Another approach computed the final impression as the averaged sum of the individual traits. A more recent framework considered impression formation as a dynamic process affected by the observer's motivation and attentional resources. In our work we took inspiration from this dynamic model since our goal was to adapt the impression of the agent in real-time during the interaction, and we were interested in the global impression formed by the user. Concerning impression management, researchers identified two main components affecting this process: the motivation to create one particular impression in the others, as well as the type of impression that one aims to convey and how to realise it. Among the factors that affect these two main components, some are very important for our work. For example, the main motivation of our agent when creating an impression into the user was to reduce the discrepancy between its desired image (e.g., to be perceived as warm) and the current impression of the user. In addition, user's preferences were an important factor which defined the type of the impression the agent wanted to provoke.

The key points of this Chapter:

- First impression involve two main processes: impression formation and impression management.
- Impression formation is achieved through three main steps: information processing, attribution of the causes of others' behaviour and information integration.
- Different approaches tended to explain impression formation. The main ones focused on impressions in their globality, tried to compute impressions as the averaged sum of individual traits or considered impressions as a dynamic process influenced by observer's motivation and attentional resources.
- Impression management is a function of two main factors, that are impression motivation and impression contstruction.
- In the context of impression management, non-verbal behaviour is less intentionally controllable compared to verbal behaviour.

Chapter 3

Warmth and Competence

Contents

This Chapter introduces the two fundamental dimensions of social cognition, i.e., warmth and competence, on which the work presented in this Thesis is based. Different frameworks and terminologies concerning these dimensions are presented, as well as evidence for the centrality of these variables in different domains. We then describe how they can relate to each other, according to different contexts. Finally, we outline non-verbal behaviours that can influence people's impressions of warmth and competence.

3.1 The Two Fundamental Dimensions of Social Cognition

As we have seen in the previous Chapter, during social interactions many cognitive mechanisms are involved, such as as processing, storing and applying information about other people. These activities are defined as social cognition. Under an evolutionary point of view (Fiske et al., 2007), social cognition reflects the survival need of knowing the intentions of the others (positive or negative), and the consequent ability (or failure) to enact those intentions.

Two broad classes of content (i.e., others' intentions and capabilities) are processed during the first moment of an encounter. These two dimensions have been studied by several researchers, under different points of view and using different labels (Cuddy et al., 2008). Several authors highlighted their centrality in both inter-personal (Rosenberg et al., 1968) and inter-group perception (Fiske et al., 2002), as well as the unique emotional and behavioural consequences of their judgments (Cuddy et al., 2008).

In the next paragraphs we describe the main frameworks used by different authors to identify these two classes of content. They are resumed in Table 3.1. It is to notice that many authors did not provide an explicit definition of these classes, but rather associated them with a list of traits.

Self-profitable vs other-profitable traits.

Peeters (1983) distinguished traits according to two perspectives: self-perception and other-perception. *Self-profitable* traits are those attributes that directly benefit or harm the actor who performs the act in question, thus allowing him to effectively pursue his goals. *Other-profitable* traits are those attributes that directly benefit or harm the observer of the act in question, informing him about other's intentions.

Social vs intellectual traits.

Rosenberg et al. (1968) identified two main categories of personality traits, which they labeled as *intellectual good/bad* (determined, skillful, industrious, intelligent, scientific) vs *social good/bad* (warm, honest, helpful, good-natured, sincere, tolerant) For more details, see subsection 3.2.1.

Morality vs competence.

Wojciszke (1994) in its Oriented Goal Theory, focused on other's intended goal, determining other's *morality*, and the efficiency of the goal attainment, determining other's *competence*. Four possible actions result from the combinations of moral and competent goals: virtuous success, when a moral goal is attained; virtuous failure, when the goal is moral but not attained; sinful success, when an immoral goal is attained; sinful failure, when the goal is immoral and not attained. These 4 actions are judged differently by observers: they are liked and respected, liked and disrespected, disliked and respected and disliked and disrespected, respectively.

Warmth vs competence.

Several authors adopted these terms to identify the two fundamental dimensions of social cognition. Cuddy et al. (2008) used a *warmth* scale including traits as good-natured, trustworthy, tolerant, friendly, and sincere; and a *competence* scale including traits as capable, skillful, intelligent, and confident. Competence entails the possession of skills, talents, and capability, but it can take the form of potential action as well as actual action. Fiske et al. (2002) used the same terms when describing inter-groups stereotypes (see subsection 3.2.2.1).

Dual Perspective Model of Agency and Communion.
More recently the *agency* and *communion* framework has been proposed. These two terms had already been introduced in philosophical context by Bakan (1966) as the basic modalities of human existence: agency as the existence of the organism as an individual, and communion as the existence of the individual belonging to some larger organism. Abele and Wojciszke (2014), in their Dual Perspective Model of Agency and Communion (DPM-AC), used the same terms to distinguish two broad classes of content universally present in the perception of the self, other persons, and social groups. These terms were: *agentic* content, which refers to goal achievement and task functioning (competence, assertiveness, decisiveness), and *communal* content, which refers to the maintenance of relationships and social functioning (helpfulness, benevolence, trustworthiness). The DPM-AC supported the hypotheses of the primacy of communal over agentic content (evidence about it can be found in subsection 3.3.1), and that communal content receives more weight than agentic content in other-perception, while the opposite occurs in self-perception.

All the different terminologies used by several authors do not substantially differ in meaning, as demonstrated by Abele and Wojciszke (2013). They found that two factors (accounting for almost 90% of the variance) can resume traits representing communion, collectivism, femininity, morality, other-interest vs traits representing agency, individualism, masculinity, competence, self-interest. Beyond different frameworks, this two broad categories are important in social cognition under different points of view. Ontologically, they reflect the humans' existential needs "to gain social acceptance and establish supportive social connection with others", as well as "to attain competencies and status" (Ybarra et al., 2008; Abele and Wojciszke, 2013). This position considers the two dimensions as the motives of humans' behaviours, and thus the two classes on which social cognition should be based. Under a functional and evolutionary account, these two dimensions reflect humans' need to determine other's intentions (beneficial or harmful), as well as other's ability to enact these intentions.

In this Thesis the terms **warmth** and **competence** (W&C) are used since they are the most used in literature about human-human and human-agent interaction: the former includes traits like friendliness, trustworthiness, sociability; the latter includes traits like intelligence, agency and efficacy.

First class of content	Second class of content	Reference
Other-profitable traits: attributes that directly benefit or harm the observer of the act in question, informing him about other's intentions.	*Self-profitable traits*: attributes that directly benefit or harm the actor who performs the act in question, thus allowing him to effectively pursue his goals.	Peeters (1983).
Social good/bad traits: warm, honest, helpful, good-natured, sincere, tolerant.	*Intellectual good/bad*: determined, skillful, industrious, intelligent, scientific.	Rosenberg et al. (1968).
Morality: it is determined by other's intended goal.	*Competence*: it is determined by the efficiency of the goal attainment.	Wojciszke (1994).
Warmth: it is measured by a scale including traits as good-natured, trustworthy, tolerant, friendly, and sincere.	*Competence*: it is measured by a scale including traits as capable, skillful, intelligent, and confident.	Cuddy et al. (2008) and Fiske et al. (2002).
Communion: it refers to the maintenance of relationships and social functioning (helpfulness, benevolence, trustworthiness).	*Agency*: it refers to goal achievement and task functioning (competence, assertiveness, decisiveness).	Abele and Wojciszke (2014).

Table 3.1 – Different frameworks used by authors to describe the two classes of content processed in social cognition.

3.1.1 Sub-components of Warmth.

Brambilla et al. (2011) suggested to split the warmth dimension into separate aspects: *sociability* and *morality*. Sociability concerns attributes linked to cooperation, such as friendliness and likeability, while morality concerns aspects pertaining to the correctness of social targets, such as honesty, sincerity, and trustworthiness. Previous studies demonstrated a different role of these 2 aspects at both inter-personal and inter-group perception (Leach et al., 2007). Brambilla et al. (2011) investigated the role of these 2 sub-components in the process of information gathering when forming impressions. To do this, they investigated which traits are primarily selected when forming impressions about others: participants were asked to evaluate the relevance of selecting certain traits from a list of 15 (5 linked to morality, 5 to sociability and 5 to competence) to accomplish four goals (global impression, morality-relevant, sociability-relevant and competence-relevant goals). Participants considered the morality-related traits more relevant than the others, independently of the goal assignment. Moreover, in a second study, they found that participants used a different strategy in searching for information on different traits. Further studies (Brambilla et al., 2012) gave additional evidence that the sociability and morality components of warmth are processed differently in information gathering.

3.1.2 Sub-components of Competence.

Similarly to the warmth dimension, the competence dimension has been proposed to be splitted into 2 aspects: *agency* and *competence*. Indeed, according to Carrier et al. (2014), the concept of competence as described by Fiske et al. (2007) does not overlap with the concept of agency adopted by Abele and Wojciszke (2013). The first referred to the capability to achieve one's intention (the nature of this intention is the warmth dimension). The second referred to a tendency to benefit the self (e.g., ambition), including an intention that does not concern the competence dimension. Moreover, also the traits mostly used in the studies relative to competence vs agency differed: competent, capable, intelligent and efficient for competence; a higher variety of traits, including active, self-confident, dominant, independent were used for agency. Carrier et al. (2014) asked participants to judge a high-status or a low-status target by choosing among agency-related, competence-related and warmth-related traits. A factorial analysis identified 3 factors, corresponding to agency, competence and warmth traits. They also identified two different relationships with warmth: agency judgments were negatively correlated with warmth, while no correlation was found between competence judgments and warmth. This distinction is in line with the different findings of Fiske et al. (2002) and Abele and Wojciszke (2013): the first argued for an orthogonality between the competence sub-component and warmth, while the others found a negative correlation between agency and warmth.

Another interesting division of the competence dimension distinguished 3 aspects according to the context application (Le Deist and Winterton, 2005). We can consider cognitive competence (knowledge, abstract intelligence and experience), functional competence (skills, accuracy and speed in performing a task) and social competence (relational and behavioural skills, the ability of managing an interaction).

3.2 W&C Dimensions in Different Domains

The centrality of warmth and competence dimensions emerged from different domains such as personality perception and group stereotypes. Here we describe into details the most important studies which contributed to identify the centrality of W&C in impression formation.

3.2.1 Interpersonal Perception

Asch (1946) was the first to intuit the centrality of the warm/cold variable in impression formation (see Section 2.1): in his studies he found for example that the term "warm" versus "cold" could shape people's "Gestalt impressions" (i.e., the general, entire impression) of a fictitious person described by a list of competence-related characteristics, while other peripheral qualities, such as "polite" versus "blunt", did not produce the same effect. Kelley (1950) extended this effect to the perception of a real person: he manipulated the same variable in a task where subjects were given preinformation about the person to judge. In this study the target person was real and interacted with the subjects. Participants given the "warm" preinformation consistently rated the stimulus person more favorably than did those given the "cold" preinformation. In addition, the study showed an impact of the variable on the interaction, with subjects given the "warm" preinformation interacting more than the others.

Rosenberg et al. (1968) quantitatively demonstrated the centrality of both W&C traits in personality impressions: they ask to sort 64 traits into categories and used multiple-regression techniques to estimate their position in the trait space. More specifically, participants were asked to think of a number of people they knew, who were different from one another, and choose the traits among the list which best described each individual. Two dimensions were found to best represent the general trait structures of person judgments: one axis for social desirability (good vs bad social traits) and one for intellectual desirability (good vs bad intellectual traits).

Another quantitative evidence for the centrality of W&C dimensions in interpersonal perception was given by Wojciszke and colleagues' studies. In Wojciszke (1994), participants were asked to think about real episodes that led them to form a clear evaluation of a person or themselves. The three-quarters of more than 1000 evaluations were based on W&C content. Similarly, in Wojciszke and Abele (2008), when asking participants to

rate 20 well-known people on W&C traits and to give an additional global evaluation, they showed that traits ratings accounted for around 82% of the variance in global impressions. This supported the thesis that when people interpret behaviours or their impressions of others, W&C form basic dimensions that almost entirely account for how people characterise others.

3.2.2 Group Stereotypes

Numerous in-depth analyses of stereotypes of specific social groups also revealed W&C as central dimensions in inter-group perception. Two main works showed evidence of W&C role on the three components of inter-group bias, that are, stereotypes (the cognitive component), emotional prejudices (the affective component) and discrimination (behavioural component). These works are the Stereotype Content Model (SCM) of Fiske et al. (2002) and the BIAS map of Cuddy et al. (2008). According to this 2 frameworks, different combinations of W&C judgments generate different group stereotypes, as well as unique emotional and behavioural responses.

3.2.2.1 Stereotype Content Model

The Stereotype Content Model of Fiske et al. (2002) attempted to explain the content and characteristics of groups stereotypes, resumed in Figure 3.1. The authors supported the universality of W&C dimensions in social perception, they investigated the social structural origins of these dimensions and the consequences of judgments about them on emotional prejudices. The main tenets of the model stated that:

- *Group stereotypes vary along W&C dimensions*: similarly to interpersonal perception, people evaluate others based on their intentions towards the in-group (to harm vs to facilitate it) and their capability to pursue these intentions (active vs passive). Groups stereotypes qualitatively differ according to the potential impact of the out-group on the in-group, that is, their perceived W&C.

  Fiske et al. (2002) conducted several studies to investigate this hypothesis. They started with a pilot study aimed to obtain a list of groups people spontaneously use to classify other. Then, in Studies 1-3 of Fiske et al. (2002), they asked participants to evaluate these groups on W&C traits. Cluster analyses consistently revealed clusters of groups that fitted specific warmth–competence combinations. In particular, four stable clusters consistently accounted for 70% of the groups, across different samples.

- *Many stereotypes fall in ambivalent but functionally consistent combinations*, where one dimension is positively evaluated and the other is negatively evaluated. Paternalistic stereotypes concern groups viewed as warm and non competent (e.g., elderly people, speakers of nonstandard dialects, housewives), while envious stereotypes

concern groups that are viewed as competent and not warm (e.g., career women, Jews, Asians). These mixed stereotypes are functionally consistent as they maintain the status quo and defend the position of societal reference groups.

To test the frequency of mixed combinations, Fiske et al. (2002) in their studies 1-3 examined the distribution of groups into various clusters and assessed differences in W&C ratings for each group. Roughly half the groups showed consistently mixed stereotypes across samples and methods of analysis.

- *Status and competitions predict dimensions of stereotypes*: when analysing the relationship between social structural variables, such as status and competition, and W&C, perceived competition was found to be highly correlated with perceived competence and perceived competition was negatively correlated with lack of warmth.

- *Different combinations of stereotypical W&C result in unique inter-group emotional responses (prejudices)*. Paternalistic stereotypes (portraying non competent and warm groups) elicit pity and sympathy, while envious stereotypes (portraying competent and not warm groups) elicit envy and jealousy. The other combinations of W&C include contemptuous stereotypes, eliciting anger and disgust, for not competent and not warm groups, while in-groups and reference groups, viewed as competent and warm, elicit admiration and respect.

  This hypothesis was supported by study 4 in Fiske et al. (2002), where participants were asked to rate the same social groups used in the other studies on emotions traits. Cluster analysis revealed that emotions scores differed significantly within all clusters and fit the emotional outcomes predicted by the SCM model.

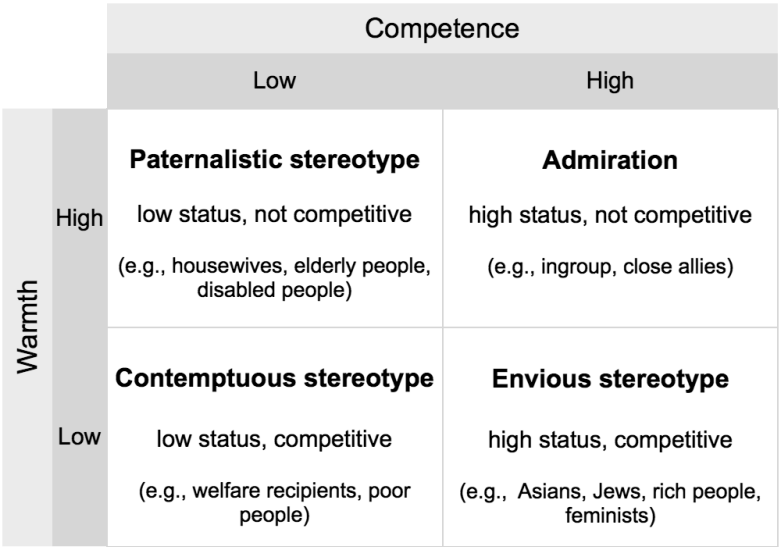


Figure 3.1 – Stereotype Content Model, adapted from Fiske et al. (2002): four types of stereotypes resulting from combinations of perceived W&C.

3.2.2.2 Behaviours from Intergroup Affect and Stereotypes (BIAS) Map

Cuddy et al. (2008) proposed another framework extending SCM by considering the behavioural outcomes of W&C judgements and inter-group responses in social interaction. Two axes were proposed to predict the behavioural tendencies linked to different stereotypes: active-passive dimension, concerning the intensity and effort of the behaviour; harm-facilitation, concerning the valence, the intended effect of the behaviour. Active and passive behaviours both affect the target group, with active ones acting with more directed effort than do the passive ones. Facilitation aims to benefit the target group, while harm acts against the target group. The combinations of these 2 dimensions lead to 4 different behavioural tendencies, predicted by W&C stereotypes and their corresponding social emotions. These combinations are depicted in Figure 3.2.

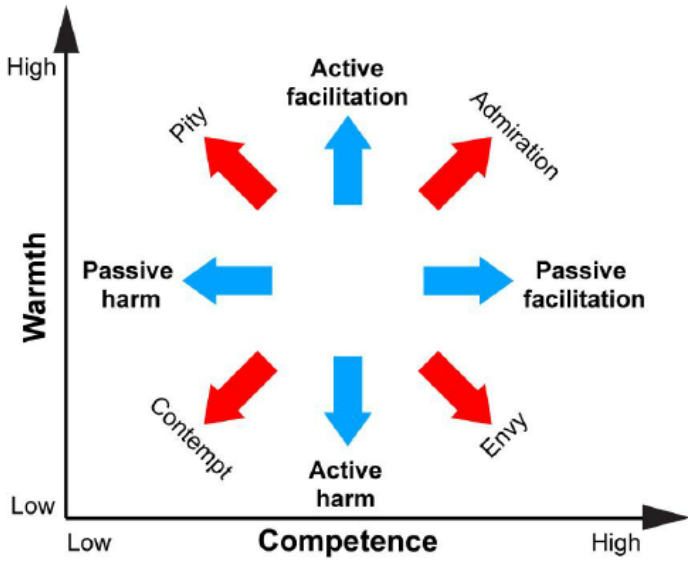


Figure 3.2 – BIAS map: schematic representation of behaviours from inter-group affect and stereotypes. Warmth and competence stereotypes are represented along the *x*- and *y*- axes. The red arrows represent emotions and the blue arrows represent behavioural tendencies (Cuddy et al., 2008).

The experimental studies conducted by Cuddy et al. (2008) supported the 3 hypotheses of the BIAS model:

- *Stereotypes predict behaviours.* In particular, the warmth dimension is supposed to predict the valence of active behaviours, while competence is supposed to predict the valence of passive behaviours. This prediction is based on the evidence for the primacy of warmth information over competence (see Section 3.3) and it is in line with the perspective that warmth is a positive response to others, actively conveyed (Bayes, 1972). For example, people who are not warm are more dangerous than not competent people, thus they elicit more urgent reactions (in this case, active harm). In particular, the behavioural tendencies predicted by warmth stereotypes

include helping others (active facilitation) and preventing behaviours such as attacking others (active harm), while those predicted by competence stereotypes include convenient cooperation, acting with others (passive facilitation) and preventing behaviours such as neglecting and excluding others (passive harm).

The correlation studies conducted by Cuddy et al. (2008) supported this hypothesis. Indeed, warmth ratings about social groups correlated positively with active facilitation and negatively with active harm. Competence ratings correlated positively with passive facilitation and negatively with passive harm. Groups stereotyped as possessing warmth elicited more active facilitation and less active harm than groups stereotyped as lacking warmth. Groups stereotyped as competent elicited more passive facilitation and less passive harm than groups stereotyped as lacking competence.

- *Emotional responses predict behaviours.* According to this hypothesis, admired groups, coming from high W&C stereotypes, lead to facilitation (active or passive). Envied groups, coming from high competence and low warmth, elicit both resentment and respect, and could lead to both passive facilitation and active harm. Contempt groups, coming from low W&C stereotypes, lead to harm (active or passive). Pitied groups, coming from high warmth and low competence stereotypes, include both compassion and sadness, and could lead to both active facilitation and passive harm.

  Correlational data in the study conducted by Cuddy et al. (2008) strongly supported seven of eight of the specific predicted links.

- *Emotions more strongly and directly predict behavioural tendencies than stereotypes.* This hypothesis relies on appraisal theories of emotions, arguing for a stronger relation between emotion and behaviour, compared to cognition.

  Cuddy et al. (2008) in their studies found that admiration fully mediated the relationship between warmth stereotypes and active facilitation, and partially mediated the relationship between competence stereotypes and passive facilitation. Contempt fully mediated the relationship between warmth and active harm, and pity fully mediated the relationship between competence stereotypes and passive harm. Only envy did not mediate any relationships of stereotypes to behavioural tendencies.

3.3 Asymmetrical Diagnosticity

Information processing of positive vs negative W&C traits is asymmetrical: evidence showed that negative warmth-related information has stronger effect than positive information, while the opposite occurs for competence-related information. Many studies showed how people give more importance to information confirming other's competence and disconfirming other's warmth (among others, Skowronski and Carlston (1987); Singh and Teoh

(2000)). According to Reeder et al. (2002) and Skowronski and Carlston (1987) explanations, positive warmth-related behaviours, being highly controllable, are not diagnostic: negative deviations are attributed to person's disposition, while positive deviations do not necessarily reflect traits but they can occur due to situational demands and reward pressure. In other words, a mean person can sometimes behave positively, but this will not change observer's global impressions. On the other hand, negative competence-related behaviours, since they are not considered under other's direct control, are less diagnostic than the positive ones. Negative deviations can be explained by a lack of motivation or the difficulty of the task, while positive deviations are attributed to other's abilities.

3.3.1 Primacy of Warmth

The asymmetrical diagnosticity of W&C in information processing suggests a priority role of warmth-related information. More evidence exists for the primacy of warmth judgments over competence ones, coming from theoretical explanations, as well as experimental evidence. Warmth is judged before competence, and carries more weight in affective and behavioural reactions.

When perceiving others, we select information to form our impressions, usually by preferring warmth information over competence (unless particular cases, such as in employment decisions). This priority is consistent with the idea that competence is linked to self-profitability and warmth to other-profitability (Peeters, 1983). According to this perspective, a lack of warmth in the others affects more the observer than does a lack of competence, since the latter does not interfere with observer's goals. Other's competence can affect observer's goal only if coupled with harmful (not warm) intentions. Accordingly, following an evolutionary explanation, other's intentions matter more to survival whether the other can act on those goals. Competence is indeed self-profitable since it is more directly rewarding or detrimental for the actor rather than to others.

A first empirical evidence for primacy of warmth judgments came from Asch's experiment (see subsection 3.2.1), where the manipulation of only one warmth-related trait in a set of competence-related traits affected the overall impression of the target person.

In another study (Wojciszke et al., 1998), among the 10 most accessible others-descriptors, selected by participants, most of them were related to warmth (only 2 are clearly related to competence). This dominance of warmth information over competence also emerged when people had to select traits they were more interested to be informed about in order to form a global impression of others. Without having a specific goal, people aimed to gather information about warmth-traits, while only in case of a competence-related goal (e.g., select a person for a role of negotiator) people tended to gather a greater amount of competence-related information. Also in the study described in subsection 3.2.1, where W&C accounted for the 82% of the variance, the 53% of this 82% was predicted by warmth traits, that resulted as a stronger predictor compared to competence

traits (Wojciszke, 2005). In addition, warmth judgments were strong and stable, while evaluations based on competence information were weaker and depended on warmth information (Wojciszke et al., 1998). These results show that warmth judgments can shape global impressions of others. De Bruin and Van Lange (1999) showed how impressions, expectations of other's cooperation, and own cooperative behaviour in a social dilemma are more strongly influenced by warmth information than by competence information. In addition, people expressed greater confidence in expectations based on warmth rather than competence information, suggesting that warmth dimension is more relevant to behaviour in this situation. Finally, warmth information was recalled better than competence information.

Cognitive evidence showed that information about warmth dimension is more cognitively accessible, more predictive, more heavily weighted in evaluative judgments. In lexical decision tasks, social perceivers identified warmth-related trait words faster than competence-related trait words (Ybarra et al., 2001; Abele and Bruckmüller, 2011; de Lemus et al., 2013). Warmth-related trait words were also categorized more quickly for valence (whether they are positive or negative) than competence-related trait words (Abele and Bruckmüller, 2011). In rapidly judging faces at 100ms exposure times, participants judged trustworthiness (that is a trait related to warmth (Fiske et al., 2007)) most reliably, followed by competence (Willis and Todorov, 2006).

3.4 Relation between W&C

Many authors investigated the presence and the nature of a relationship between W&C judgments, without reaching an agreement about it. The different results of their studies are not necessarily contradictory, but rather could be due to the different contexts and protocols adopted.

3.4.1 Orthogonal Relationship

SCM model (Fiske et al., 2002) and Oriented Goal Theory (Wojciszke, 1994) both argued for an independent relationship between W&C, that is, judgments about one dimension do not influence those about the other dimension. Indeed, in both frameworks all the combinations in the two-dimensional space existed. In inter-group perception, it is the case of the existence of both mixed (positive on one dimension, negative on the other) and univalent (positive or negative on both dimensions) stereotypes (see Figure 3.3). In interpersonal perception, according to Wojciszke (1994) perspective, both moral and immoral actions can be successful, but both types of goals can also remain unattained, thus showing the other's incompetence. This interdependence between the two dimensions generates 4 equally-distributed clusters.

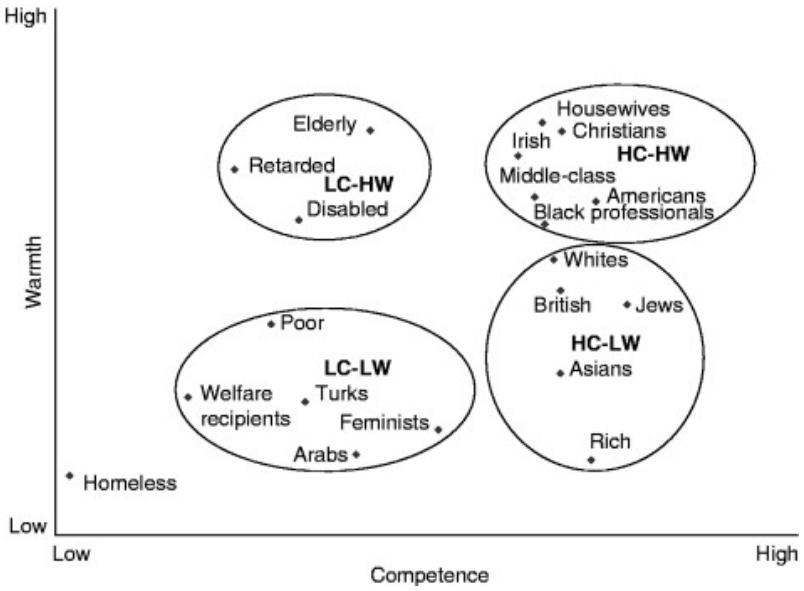


Figure 3.3 – Distribution of social groups on the W&C dimensions in the SCM (Fiske et al., 2002).

3.4.2 Positive Relationship: Halo Effect

When looking at Rosenberg et al. (1968) work, as we can see in Figure 3.4, the social good-bad and intellectual good-bad dimensions coming from their analyses are not orthogonal, but rather positively correlated. The angle between the 2 axes measured 65 degrees, corresponding to a positive correlation of 0.42. This means that positive judgments about one dimension positively affect those about the other one, in line with the so called *halo effect*. This effect was previously defined by Thorndike (1920) as the tendency to "think of a person in general as rather good or rather inferior and to color the judgment of the separate qualities by this feeling".

3.4.3 Negative Relationship: Compensation Effect

As opposed to Rosenberg et al. (1968) work, a large amount of studies revealed a negative relationship between W&C judgments. The first who coined the term *compensation effect* was Yzerbyt, who, with his colleagues, in Yzerbyt et al. (2005), conducted a set of studies to examine inter-group characterization in the context of national groups stereotypes. Participants (French vs Belgians) were asked to indicate their perception of their in-group and the out-group in terms of linguistic skills, competence and warmth. They were also asked to rate their impressions of the way they thought their own group was being seen by members of the out-group ("meta-stereotypes"). A compensation pattern emerged from their answers (see Figure 3.5): Belgians generally received lower ratings than the French on linguistic skills (from both groups) and competence, and higher ratings on warmth. They also thought to be seen by French as being not only less linguistically skilled but also

Figure 3.4 – Personality traits on the two dimensions of *social good–bad* and *intellectual good–bad* (Rosenberg et al., 1968). The two axes are not orthogonal but positively oriented.

less competent than warm. It seems like both groups tended to compensate their negative reputation in one dimension by judging them more positively on the other dimension.

Judd et al. (2005) further investigated the compensation effect in different contexts in the absence of stereotypical beliefs. They conducted several studies involving 2 artificial groups, the *Green* and the *Blue*, of which they manipulated one dimension without affecting the other one. They described the behaviours of 2 groups, one high in warmth (or competence), the other low in that dimension, and asked participants to give ratings on both dimensions. The high-warmth (resp., competence) group was judged to be less competent (resp., warm) than the low-warmth (resp., competence) group. In addition, the larger the perceived difference between the two groups on the manipulated dimension, the larger the perceived difference between them on the other dimension, in the opposite direction. From a series of follow-up studies, they found that the compensation effect occurred only in case of a 2-target comparison, that is, when participants were asked to judge 2 groups or 2 people, while halo effect was found when participants were asked to judge individual targets, outside a comparison context. This could explain the findings of Rosenberg et al. (1968).

Yzerbyt et al. (2008) investigated whether the compensation effect is just a pattern occurring in 2-target comparisons, or it is specific to W&C dimensions. To do this, they

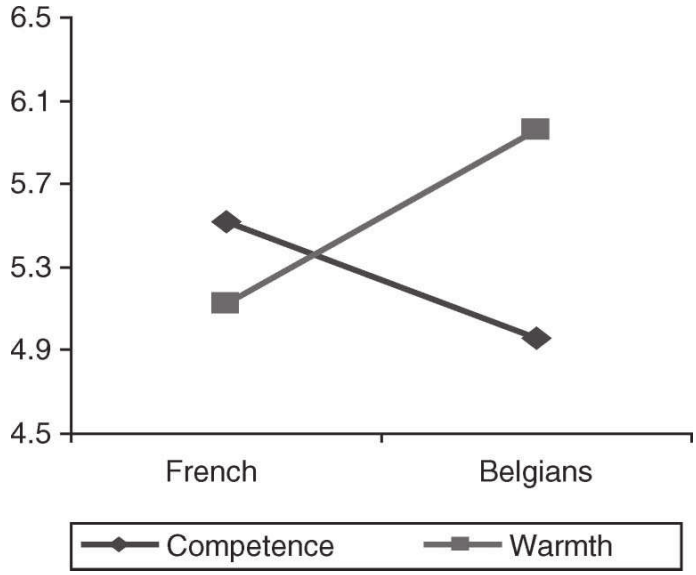


Figure 3.5 – The first occurrence of *compensation effect*: W&C perception of French and Belgians (Yzerbyt et al., 2005).

replicated the experiment of Judd et al. (2005) by manipulating W&C and asked to also rate another unrelated and unmanipulated dimension, such as healthiness. A halo effect was found for healthiness judgments, supporting the hypothesis that compensation effect is specific for W&C dimensions.

The *compensation effect* can thus be defined as the "tendency to differentiate two social targets in a comparative context on the two fundamental dimensions by contrasting them in a compensatory direction" (Kervyn et al., 2009).

Kervyn et al. (2010) proposed 2 possible interpretations of the origin of this tendency. On one hand, it could reflect an adherence to mixed stereotypes (Reinhard et al., 2008): social perceivers could have pre-conceived stereotypes about social targets to manifest a compensated pattern, that is, to be either warm and not competent, or cold and competent. Another interpretation sees the *compensation effect* as an effect of the system justification theory (Jost and Hunyady, 2002): people prefer an evaluatively balanced view of social groups in order to justify the existing social structure. When given the description of 2 groups with opposite levels in one dimension, participants react to this unjust system by compensating on the unmanipulated dimension. They bias their perception of the two groups on the dimension that was left ambiguous in order to create a system in which both groups have strengths and weaknesses, a situation that is closer to one in which both groups would have an equal amount of positive characteristics.

3.4.4 Innuendo Effect

Kervyn et al. (2012) showed evidence for the presence of a negative relationship between W&C even outside of a comparative context. They defined the *innuendo effect* as the

"tendency for individuals to draw negative inferences from positive descriptions that omit one of the two fundamental dimensions of social perception". Their findings suggest that halo, compensation and innuendo effects do not exclude each other and they are used to construct, maintain, and convey impressions consistent with social norms and stereotypes.

3.5 Non-verbal cues of W&C

Impressions about others' W&C are not only obtained from observation or description of overt behaviours (such as actions), like in most of the studies presented in this Chapter, but they can also be elicited by particular non-verbal cues. In this section we present research about non-verbal behaviours related to the expression of different levels of W&C.

Bayes (1972) attempted to define and specify the behavioural cues of warmth, by searching for an association between global ratings of warmth and objective measures of specific behavioural cues. In his experiment, videos were recorded of interviewees responding to questions about topics like home, school and job. The first 3 minutes of these videos were judged by participants on a warmth scale. One group rated the video without speech content, a second group rated the audio only and a third group rated the videos with audio. No definition of warmth was given to the raters. Participants were also asked to indicate the behavioural cues that they used to rate the interviewees. These cues included posture, head movements, hand movements, facial expressions and smiling. This last cue was found to be the best single predictor of warmth.

Cuddy et al. (2008) studied the non-verbal cues conveying W&C. For warmth, they cited the role of Duchenne's smile (Duchenne, 1990), the presence of immediacy cues that indicate positive interest or engagement (e.g., leaning forward, nodding, orienting the body toward the other), touching and postural openness, mirroring (i.e., copying the non-verbal behaviours of the interaction partner). For coldness, they cited tense posture, leaning backwards, orientating the body away from the other, tense and intrusive hand gestures (e.g., pointing). Concerning competence, they cited non-verbal behaviour related to dominance and power, such as expansive (i.e., taking up more space) and open (i.e., keeping limbs open and not touching the torso) postures. People who express high-power or assertive non-verbal behaviours are perceived as more skillful, capable, and competent than people expressing low-power or passive non-verbal behaviours.

An interesting study on the effect of hand gestures on social perception (Maricchiolo et al., 2009) showed significant effects of hand gestures type on competence perception. In their experiment, researchers created 5 different videos where an actor was performing the role of a University delegated discussing about the University Council decision to increase tuition fees. The videos differed only in the type of gestures performed by the actor: ideationals (that is, gestures related to the semantic content of the speech), beats (rhythmic gestures, linked to the speech structure and rhythm), object-adaptors (hand movement of contact with objects), self-adaptors (hand movement of contact with

parts of one's own body) and absence of gesture. Participants were asked to evaluate the speaker on a list of items, the communicative style of the speaker, the message persuasiveness, their attitude (favourable or unfavourable) towards the fees increase, and their intention to vote about it. Ideationals and object-adaptors resulted in a higher level of competence judgments, compared to absence of gestures, while self-adaptors resulted in a lower competence. No significant effect of hand gestures was found for warmth.

This last study is one of the few works that gave insights about the role of communicative gestures in conveying different impressions of W&C. In this Thesis we wanted to further investigate non only the different roles of the type of gestures (ideationals, beats, adaptors) but also the role of arms rest poses. Few works investigated this, in particular the association between rest poses and dominance, but not explicitly with W&C impressions. The work presented in Chapter 5 was conducted for this purpose. We investigated the association between W&C impressions and non-verbal behaviour like type of gestures, arms rest poses, smiling behaviour and head rotation.

3.6 Conclusion

WHEN we meet new people, we rapidily collect information about their intention towards us, as well as their capability to attain this intention. We use the same criteria when forming stereotypes about out-groups. These two broad dimensions have been called with different labels, that overlap in the meaning. We chose to use the terms warmth and competence: the former includes traits like friendliness, trustworthiness, sociability; the latter includes traits like intelligence, agency and efficacy. Warmth and competence are central in inter-personal and inter-group perception and they elicit unique emotional and behavioural outcomes. Evidence supports the primacy of warmth-related judgments over competence, in line with the other-profitability that characterizes warmth. No agreement exists about the type of relationship between the two dimensions: they are interdependent according to some frameworks, while in general they are positively correlated in a single-target context, and negatively correlated in a two-target comparison context. Our impressions about others' warmth and competence are not only obtained from observation or description of overt behaviours (such as actions), but they can be elicited by particular non-verbal cues, such as open vs closed postures, types of gestures and smiling.

The key points of this Chapter:

- Warmth and competence are the fundamental dimensions of social cognition, at both inter-personal and inter-group level.
- Warmth-related judgments come before competence-related judgments.
- Different types of relationship have been found between warmth and competence: interdependence, halo effect, compensation effect.
- Non-verbal behaviour can affect warmth and competence impressions.

Part III

Related Work

Chapter 4

Related Work

Contents

In this Chapter we provide a review of the main works addressing the topic of interest of this Thesis. At first, we present studies focusing on the impact of several variables, such as turn-taking behaviour, non-verbal cues and appearance on user's impressions. We then describe the studies that included warmth and competence dimensions in a virtual agent. Field studies conducted in public spaces are then reviewed, as well as different methods to detect and foster engagement in human-agent interaction. Finally, we discuss how our work was inspired and at the same time differed from those presented in the Chapter.

4.1 User's Impressions in Embodied Conversational Agents

Several researchers conducted evaluation studies assessing user's impressions of an agent's friendliness, dominance, agreeableness and extraversion. The emphasis has been mainly on agent's characteristics such as interpersonal attitude towards the user, personality and skill level in a selected context (e.g. competence), that can be extrapolated from brief observations of multi-modal behaviour.

4.1.1 Turn-taking Behaviour

Some studies focused on the impact of turn-taking behaviour on user's impressions of an agent. Ter Maat et al. (2010) showed how a realization of a simple communicative function (for managing the interaction) could influence users' impressions of personality (agreeableness), emotions and social attitudes (i.e. friendliness). They simulated a conversational interviewing agent using different turn-taking strategies (i.e. the management of when to speak). These strategies differed according to the duration of the pause that the human wizard controlling the agent waited after the end of the participant's speaking turn, before starting to speak. They also manipulated agent's strategy, gender and the topic of the conversation. Results revealed that starting the speaking turn too early (that is, interrupting) was mostly associated with negative and strong personality attributes, while leaving pauses between turns made the agent more agreeable, less assertive, and created the feeling of having more rapport.

4.1.2 Non-verbal Behaviour

Other studies focused on the impact of non-verbal behaviour on user's impressions.

Fukayama et al. (2002) proposed and evaluated a gaze movement model that enabled a virtual agent to convey different impression on users, in terms of affiliation (friendliness, warmth) and status (dominance, assurance).

Cafaro et al. (2016) developed a model of first impressions in user-agent interactions. In three empirical studies they manipulated an agent's non-verbal immediacy cues (i.e. smile, gaze and proxemic behaviour) during the initial phases of a greeting encounter in a museum. These manipulations were done in order to affect users' impressions of the agent's personality and interpersonal attitudes. They conducted three experiments:

1. A non-verbal behaviour interpretation study (Cafaro et al., 2012). They manipulated agent's nonverbal immediacy cues of smile (no vs. yes), percentage of gaze towards the user (low % vs. high %) and proxemics (no step towards the user vs. step) of a greeting agent in a first virtual encounter. Participants, represented by an on-screen avatar, approached a series of greeting agents in a 3D virtual museum entrance. Their impressions of agent's extraversion and affiliation were measured. Proxemics

behaviour affected user's impressions of agent's extraversion, whereas smile and gaze had significant impact on the perception of friendliness. Smiling agents and gazing agents were considered more approachable and likable.

2. A non-verbal behaviour impact study (Cafaro et al., 2013). In this study they investigated the effect of agent's personality and attitude impressions on users' willingness to interact with an agent again. Participants observed a series of animated views of first-person perspective approaches to life-sized agents presented as guides of a virtual museum. Agent's personality and attitude were manipulated according to the results of the previous study. After meeting each guide, participants were immediately asked to express, in general, "how likely they were to spend time with it again" and "their preferences about the number of guided tours they would be willing to take with the agent". Results showed that high friendliness agent received a higher number of visit preferences than low friendliness one. In addition, participants were more likely to meet them again later, and they were the most preferred ones regardless of the level of personality that was exhibited.

3. A public space study (Cafaro et al., 2016). This study aimed to test the effectiveness of managing first impressions for an agent installed in a real public setting (the Science Museum in Boston). They incorporated their first impression model in the relational agent Tinker (see Section 4.3). Thus, in addition to its relational capabilities, Tinker could also exhibit friendly greeting behaviours. The experiment aimed to determine whether users' impressions about Tinker had an impact on their decisions to approach and initiate interaction with the exhibit. Three versions of Tinker were compared: friendly, unfriendly, by manipulating its smile and gaze, and a control condition where it did not exhibit any nonverbal reaction towards the user. A total of 15286 users participated in the study. Results showed that the friendly agent encouraged the user to continue the interaction.

On the whole, these studies demonstrated that first impressions of personality and interpersonal attitude can be influenced by nonverbal immediacy cues of smile, gaze and proxemics during a first user-agent encounter, similarly to what occurs in human-human interaction (see Sections 2.2 and 3.5). These impressions have been found to affect users' relational decisions in terms of likelihood and frequency of further encounters.

4.1.3 Appearance

Another series of studies focused on the effects of appearance on user's impressions of pedagogical agents, defined as "virtual characters embedded in multimedia learning environment that simulate human instructional roles" (Liew et al., 2013; Johnson et al., 2000). Two types of pedagogical agents can be distinguished. Expert-like agents are characterised by advanced knowledge and competence in a subject domain (Baylor and Kim, 2005) and

can be operationalised through the image of "a professor in his/her forties, who speaks in formal, authoritative, and professional manner". Peer-like agents are socially similar to learners, and can be operationalised through the image of "a casually-dressed student in his/her twenties, who speaks in friendly and emphatic manner" (Kim, 2007; Kim and Baylor, 2006; Baylor and Kim, 2005).

Rosenberg-Kima et al. (2008) found that peer-like agents, that is, agents that resembled to participants (i.e., young and "cool" female virtual model) were effective in influencing female college learner's self-efficacy and willingness to enroll in engineering courses. Male virtual agents that resembled to prototypical engineers (i.e., expert-like agent) influenced the learners' perceived utility of engineering. The two agents used in the study are shown in Figure 4.1

Figure 4.1 – The male and female agents used in Rosenberg-Kima et al. (2008).

Liew et al. (2013) investigated whether expert-like and peer-like agents differently affect user's impressions about them and learning achievement. In their experiment, they manipulated visual appearance and voice inflection of the agents. The expert-like agent resembled a female college lecturer in her 40s, with strong and authoritative voice. The peer-like agent resembled a female college student in her 20s, with a soft and calm voice. The two agents are shown in Figure 4.2.

Participants watched a multimedia presentation about basic programming made by one of the agents and then they were asked to fill in questionnaires about their impressions of the agent and to complete a learning task. Participants perceived peer-like agent to be more friendly than expert-like agent. No effect of image/voice stereotypes of peer-like agent and expert-like agent was found on participants' impressions about agent's knowl-

Figure 4.2 – The peer-like and the expert-like agents used in Liew et al. (2013).

edgeability. An effect on agent's trustworthiness was found only for female participants: they assigned higher scores to expert-like agent than to peer-like agent.

In Veletsianos (2010), participants were asked to view a tutorial presented by one of the two pedagogical agents shown in Figure 4.3. They were identical in face shape, facial expressions, body image, clothing, dimensions, voice, and animation. They differed in hair style and colour. One agent also wore a necklace. Their difference was validated by a manipulation check. The first agent was called a "scientist" and the second was called an "artist".

Figure 4.3 – The "scientist" and the "artist" agents used in Veletsianos (2010).

The tutorial could concern nanotechnology or punk rock and could be presented by the scientist or the artist agent. The artist agent was rated as being more knowledgeable than the scientist agent. In addition, the appearance of the agent, in relation to the content area under consideration, influenced users' perceptions and learning. Participants recalled significantly more information when interacting with the artist than with the scientist agent. Additionally, they recalled more information when participating in the punk rock tutorial than in the nanotechnology tutorial. For the nanotechnology tutorial, participants assigned to the artist group recalled more content than participants assigned to the scientist group. For the punk rock tutorial, participants assigned to the artist group recalled more content than participants assigned to the scientist group.

The studies presented in this paragraph present some limits related to their methodology, indeed the characters used in the experiment were static and they only moved lips while talking. In addition, these studies showed that agent's appearance and voice can affect users' learning, but they are not enough to elicit consistent impressions of agent's likeability (i.e., warmth) and knowledgeability (i.e., competence). These results encouraged us to manipulate the non-verbal behaviours of the agent in order to affect the impression it gave to the user.

In the next Section we resume the few studies that explicitly investigated the role of W&C in ECAs.

4.2 Warmth and Competence in Embodied Conversational Agents

Niewiadomski et al. (2010) analyzed how the emotional multi-modal behaviour of a virtual assistant expressing happiness, sadness and fear influenced user's judgments of agent's warmth, competence and believability. In their experiment, they simulated a typical virtual assistant scenario. At the beginning, participants were asked to imagine that they possessed a new computer including a virtual assistant, that they were playing a game and that they lost it. Then, they watched a video of the reaction of the agent to their defeat. In this video, the agent could display socially appropriate or inappropriate (i.e., that were -or were not- expected in that situation) and plausible or implausible (i.e., that could be displayed even if not appropriate in that situation) emotional reactions, with verbal, non-verbal (facial expressions accompanied with emotional gestures) or both modalities behaviour. They found an impact of socially adapted emotion on agent's believability, competence and warmth. In particular, socially appropriate emotions expressed by the agent led to higher perceived believability, warmth and competence of the agent. Ratings of these variables also increased with the number of modalities used by the agent: judgments about these dimensions were higher when using multi-modal behaviours than one modality alone (either verbal or non-verbal). Then they also found that the perception of agents' believability was highly correlated with the two major socio-cognitive dimensions of W&C.

The results of this study are interesting since they are consistent with previous findings about W&C that we described in Chapter 3. Indeed, the positive correlation between W&C supports the presence of a *halo effect*, and the highest effect size of warmth judgment is consistent with the idea of a *primacy of warmth* judgments. It seems that same perceptual processes occur when people judge virtual agents and humans.

Finally, this study did not take into account user's socio-emotional and behavioural reactions to agent's behaviour.

Bergmann et al. (2012) studied how appearance and gestures affected the perceived W&C of virtual agents over time. This was the first study investigating the dynamics of first impressions about ECAs. The authors investigated how W&C ratings changed from a

first impression formed after a few seconds to a second impression after a longer period of human-agent interaction, depending on manipulations of agents' visual appearance and non-verbal behaviour. Participants were asked to rate their first impressions of the agent after that it briefly introduced itself. In the second phase of the experiment, the agent performed a task where it described a building with six sentences and was judged again by the participants. Visual appearance of the agent was manipulated with regard to its level of human-likeness. That is, agent's appearance could be robot-like or human-like. The non-verbal behaviour of the agent included two conditions: no gestures at all, or the presence of co-speech gestures, i.e., related to the semantic content of the speech. Agent's appearance influenced the way first impressions about warmth changed over time. Ratings about agent's warmth decreased at the second time measurement only for robot-like characters, while they did not change for the human-like character, as shown in Figure 4.4 (left).

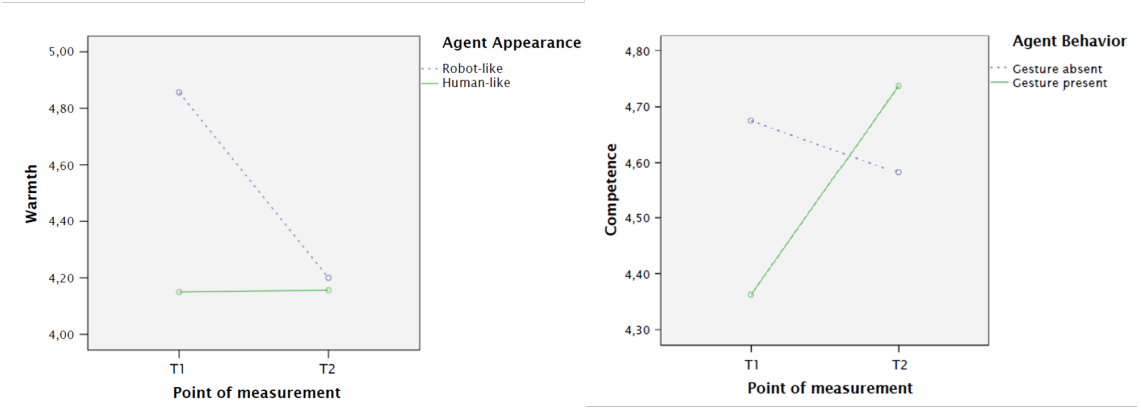


Figure 4.4 – At the left, warmth ratings as a function of agent appearance and point of measurement. At the right, competence ratings as a function of agent behaviour and point of measurement (Bergmann et al., 2012)

Concerning competence judgements, gesture presence influenced the way first impressions about them changed over time. Presence of co-speech gestures increased the perceived competence of the agent between the two points of measurement, while ratings of agent competence slightly decreased in absence of gestures, as shown in Figure 4.4 (right). These results showed that the two dimensions of social cognition were not affected by agents' behaviour and appearance in equal measures. Warmth judgments were more prone to be modified over time, in particular their dynamics were influenced by agent's appearance, while the dynamics of competence judgements seemed to depend more on gestural behaviour.

In this study only co-speech gestures were manipulated, while other non-verbal behaviours were not considered, in particular smiling behaviour. Smiles are important in conveying warmth impressions in human-human interaction (see Section 3.5), thus in our

work we were interested in investigating how a virtual agent is perceived according to different smiling behaviour.

4.2.1 A Computational Model of W&C in Virtual Agents

The studies presented so far investigated the effect of some controlled variables, like presence or absence of gestures, on users' judgements about agent's W&C, in specific scenarios.

Nguyen et al. (2015) were the first to develop a computational model for generating agent's behaviours eliciting different levels of these dimensions. They followed "an iterative design methodology tuning the design using theory from theater, animation and psychology, expert reviews, user testing and feedback". Actors were asked to perform a given text with different degrees of W&C. The gestures of the actors were analysed by experts on 4 dimensions: speed, weight, time and flow, according to Laban Movement Analysis, a framework often used in theatrical performance (Laban and Lawrence, 1979). Gesture space was analysed according to McNeill's Growth Point notation (McNeill, 1992). The experts who revised the design process were two psychologists, a theatrical performance director, a virtual human designer and an animator. Their role was to suggest a set of rules to be encoded in the virtual character algorithm under development. Videos of the virtual characters were then distributed to the experts panel for analysis and feedback. This process continued until the experts panel was unanimously satisfied with the resulting virtual characters. A validation study was conducted to evaluate the effect of the behaviour model on user perception. The main results are shown in Figure 4.5. They found that high-warmth agents were perceived as warmer than low-warmth ones, regardless of whether they were competent or not (Figure 4.5a). High-competence agents were perceived as more competent than low-competence ones, regardless of whether they were warm or not (Figure 4.5b). High-warmth characters were perceived as more competent than low-warmth characters. A significant warmth×competence interaction was found: the effects of intended warmth levels on W&C ratings differed depending on the intended level of competence. When a character was highly warm, its competence level did not affect how warm it was perceived. However, a low-warmth character was perceived as warmer if it was competent.

The findings of the studies presented in this Section are in line with the phenomena we described in Chapter 3 which characterise people judgments about others' warmth and competence. It seems that the same patterns occur when people judge virtual agents. In particular, support for *halo effect* was found by Niewiadomski et al. (2010) and Nguyen et al. (2015), support for a *primacy of warmth* judgements was found by Niewiadomski et al. (2010) and results of Bergmann et al. (2012) reflect the presence of *asymmetrical diagnosticity* of W&C perception.

These studies shown that it is possible to influence user's impressions of agent by managing its nonverbal behaviour, and that people tend to judge virtual agents by applying

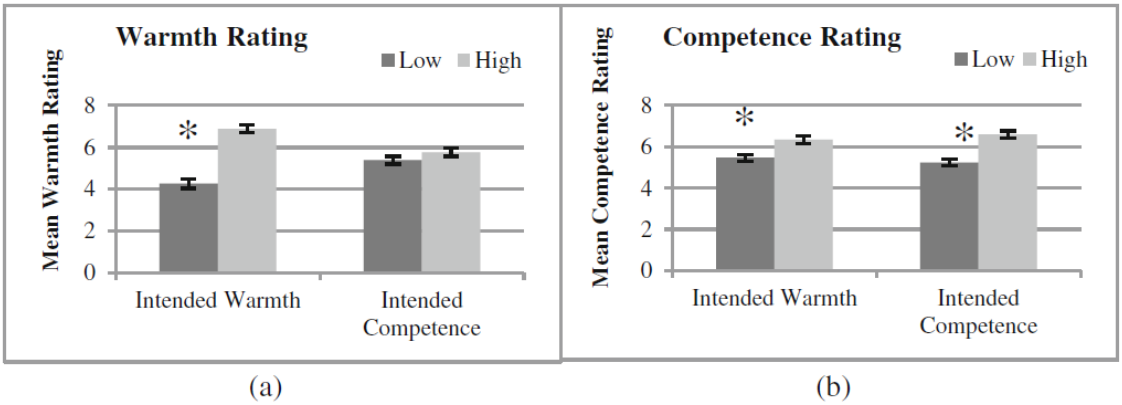


Figure 4.5 – Mean Warmth and competence means as a function of intended warmth and intended competence levels encoded in the animation videos (Nguyen et al., 2015). * stands for $p < 0.001$.

the same patterns that occur when judging humans. However, these studies present some limits. In Niewiadomski et al. (2010) the study was conducted online and the agent only performed emotional reactions, while in Bergmann et al. (2012) only co-speech gestures were manipulated. Finally, the model developed by Nguyen et al. (2015) was based on videos of actors and not natural interactions.

With respect to the studies presented in this Section, we aimed at investigating the nature of W&C impressions in human-agent interaction, with a focus on the relations between the two dimensions. With respect to Bergmann et al. (2012), we considered more behaviours than only co-speech gestures, such as facial expressions and head poses. To find what these behaviours are, we proposed a methodology that used natural interactions videos instead of videos of actors (see Chapter 5) with both discrete and continuous annotations.

4.3 Embodied Conversational Agents in Public Spaces

Most of the studies presented above, except for Cafaro et al.'s ones, were conducted online or in a laboratory setting. Virtual and robotic conversational agents have also been deployed in public spaces for field studies. These deployments allowed researchers to move from the controlled laboratory settings to a more natural but challenging real life environment which can be noisy and with the presence of multiple users. Several virtual agents were installed in museums and public halls. Here we mention the most interesting and important ones related to our research.

Gockley et al. (2005) installed the receptionist Valerie at the entrance-way to Newell-Simon Hall, in Carnegie Mellon University. They conducted a long-term interaction study aimed at investigating how a social robot can remain compelling over a long period of time. Valerie, depicted in Figure 4.6, was built on a mobile base with a moving flat-panel

monitor mounted on top, displaying a graphical expressive human-like face. It had its own personality and a personal story evolving over time, scripted by students of a Drama School. User could interact with it through a keyboard. Valerie could perform storytelling through monologues, but also give information like giving directions, looking up weather forecasts, etc. A card reader allowed users to swipe any magnetic-stripe card in order to be identified. Data from the card were stored and used to remember user information from one interaction to the next one. After nine-month period of operation, it was found that users continued to interact with Valerie on a daily basis, even after the "novelty effect" faded. Among users who used their card to be recognised, 200 people chose to interact with Valerie multiple times. Most people stayed long enough to greet the robot and hear one or two sentences of a monologue, but not more. Only visitors who came several times interacted and listened for much longer periods.

Figure 4.6 – The permanent installation of the receptionist Valerie (Gockley et al., 2005). It was built on a mobile base with a moving flat-panel monitor mounted on top, displaying a graphical expressive human-like face.

(Kopp et al., 2005) employed the conversational agent Max (Kopp and Wachsmuth, 2004) in a real-world setting. The agent played the role of a guide at the Heinz Nixdorf Museums Forum (HNF), a public computer museum in Paderborn (Germany). Max's primary task was to engage visitors in conversations, by providing them with information about the museum. Max was displayed in human-like size on a screen, standing face-to-face to the visitors of the museum. Cameras allowed it to notice visitors that were passing by. Users could interact with it through a keyboard. This input device was chosen in

order to avoid problems that could arise from speech recognition in noisy environments. Max was equipped with an emotion system that continuously ran a dynamic simulation to model the agent's emotional state. The system used a rule-based approach to dialogue modeling which took into account the context-dependent aspects of dialogue acts. Max accompanied text with fully multi-modal behaviour. The authors conducted field studies in the museum to evaluate how users interacted with Max. They analysed the content of visitors' dialogues, in order to find out whether people used human-like communication strategies when interacting with Max and whether they used utterances that indicated the attribution of sociality to the agent. They found that people were likely to use human-like communication strategies (greeting, farewell, small talk elements and sometimes even insults) and they were cooperative in answering Max's questions. This supported the attribution of social traits to the agent.

Figure 4.7 – The agent Max interacting with visitors in the Heinz-Nixdorf-Museums Forum (Kopp et al., 2005).

Bickmore et al. (2013) developed the computer animated agent Tinker, shown in Figure 4.8, which was installed in the Computer Place exhibit at the Boston Museum of Science for several years and got many positive feedback from thousands of visitors. Differently from previous examples, this was the first work which included a model of the user-agent relationship. This model included a variety of dialogues and nonverbal behaviours to enable Tinker to establish a sense of trust and rapport with visitors, with the purpose of encouraging them to continue the interaction as well as to come back and visit the museum. These social functions included: expressing empathy; asking users about themselves and making references to this information during the interaction; inserting humor at appropriate points in the conversation; expressing liking towards the user and the desire to continue the interaction. Tinker's tasks included describing the exhibits, giving directions and discussing about its own implementation. Users could interact with it

through a multiple-choice touch screen. A glass plate with sensors could detect the presence of users, who could also be identified thanks to the biometric analysis of their hand shapes. Tinker could use this information to re-identify return visitors and continue the interaction with them. The authors conducted two experimental studies to evaluate the effect of Tinker's relational behaviour on engagement and learning. The first study aimed to test the impact of automatic re-identification on the attitude of visitors towards Tinker. For this, the version of Tinker with the hand recognition was compared to another version without the biometric hand reader, so that Tinker did not recognise return visitors. No significant effects of re-identification were found, probably due to the very small number of participants who completed the entire study protocol (only 29). The second experiment aimed to evaluate the impact of relational behaviour on visitors' engagement and learning. For this, the full version of Tinker was compared to another version without the relational behaviour model. They collected data from 1607 visitors. Users' engagement, measured by session length, number of sessions, and self-reported attitudes, as well as learning, measured by a knowledge test, was higher for people who interacted with the full version of Tinker. In addition, regression analysis showed that the use of relational behaviour affected learning through increased engagement: when Tinker used relational behaviours, users spent more time interacting with Tinker and so learned more.

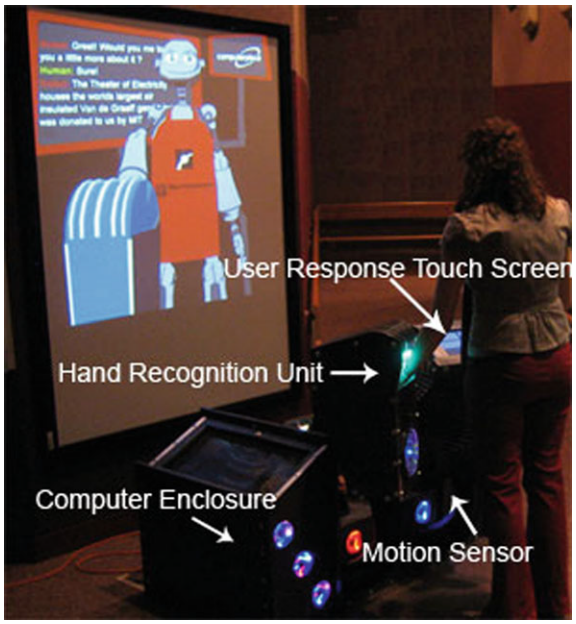


Figure 4.8 – Tinker installation (Bickmore et al., 2013). A glass plate detected the presence of the user, who could be recognised by a hand reader. The interaction with the agent was possible through a touch screen.

Another important work was realised by Swartout et al. (2010) and concerned the implementation of the Virtual Museum Guides Ada and Grace, two virtual humans set in an exhibit at the Science Museum in Boston. In figure 4.9 they are interacting with the

visitors. The system was based on natural language interaction. Ada and Grace's goal was to engage children, in particular females, into discussions about Science, Technology, Engineering and Mathematics. Authors chose to use two characters to enhance people's engagement. The presence of two characters allowed to share long responses and avoid users getting bored. In addition, the two agent could express different points of view and use humour to engage the user.

During the first year of the system's deployment, visitors' questions were mediated by a museum staff member who interacted with the characters, in order to limit problems related to speech recognition and to allow for better data that were used for training visitor models. From the second year, visitors were able to directly talk with the characters. The audio was sent to the automatic speech recognizer (ASR), which passed the transcription of user's speech to a Language Understanding (LU) module. This module selected a set of appropriate responses from a domain-specific library of scripted responses and sent them to the dialogue management (DM) module. The DM module selected the final response from the set of the possible ones by taking into account the recent dialogue history in order to avoid repetitions.

This installation allowed collecting a large database of utterances from museum visitors and museum staff, spoken in interaction with the Twins (Aggarwal et al., 2012). The corpus contained about 200,000 spoken utterances and was used for improving the dialog model deployed at the museum (for example by identifying the most common questions in order to improve agents' answers), as well as for other research project to improve natural language processing.

Figure 4.9 – Visitors interacting with Ada and Grace (Swartout et al., 2010).

In summary, field studies have been conducted to engage users in interactions with a virtual agent. These studies focused on particular dimensions of first impressions such as interpersonal attitudes and personality in order to make agents more engaging and accepted for long-term interactions. However, none of them dealt with impression management nor proposed a system integrating the assessment of user impressions, particularly detecting users' multi-modal cues for implementing the agent's behaviour adaptation mechanisms.

4.4 Engagement in Human-Agent Interaction

Engagement plays an important role in human-agent interaction: an engaging virtual agent is more likely to be accepted by the user, as well as to promote future interactions Bergmann et al. (2012), Cafaro et al. (2016).

User's engagement has been widely studied by the agent community, as shown in Clavel et al. (2016). In their review about engagement in human-agent interaction, they stated that an ECA should take into account user's social attitudes and emotions as cues of user's engagement or disengagement. There exist many model to analyse user's non-verbal and verbal emotional content (Osherenko et al., 2009; Schuller et al., 2004; Smith-Lovin and Brody, 1989; Lin, 2009; With and Kaiser, 2017; Clavel and Carrión, 2016).

On the other hand, the agent should not only be capable to detect the presence of user's state, but it should be able to react to it by expressing believable and engaging socio-emotional behaviour. Several computational models exist that allow ECAs to express different attitudes and emotions (e.g., (Chollet et al., 2013; Lee and Marsella, 2006; Ravenet et al., 2013a; Ruttkay et al., 2003; Niewiadomski et al., 2011; With and Kaiser, 2017).

4.4.1 Interaction Strategies

Several interaction strategies have been proposed in order to foster user's engagement. They focused on managing different characteristics of the agent or on applying different techniques to increase user's engagement.

4.4.1.1 Backchannels

Morency et al. (2009) used sequential probabilistic models to automatically learn from a database of human-human interactions to predict listener backchannels using the speaker multi-modal output features (e.g., prosody, spoken words and eye gaze). This model obtained a better performance for the prediction of visual backchannel cues (i.e., head nods) than other hand-crafted based rules models.

Truong et al. (2010) manually designed rules for a model predicting when to perform a bakchannel based on pitch and pause information. This model highlighted the role of the length of a pause preceding a backchannel, indeed it was found to perform slightly better than another well-known rule-based prediction model using only pitch information.

Schröder et al. (2015) created a model that decided when to trigger agent's backchannels, by applying probabilistic rules, and the type of multi-modal backchannel (smile, nod, and verbal content) to display. The agent could provide either feedback about its communicative functions (such as agreement, liking, believing, being interested and so on) (Allwood et al., 1992; Poggi, 2007) or signals of mimicry to mirror the speaker's non-verbal behaviour.

4.4.1.2 Politeness

Andre et al. (2004) integrated politeness strategies with a cognitive theory of emotions. According to their model, the agent should increase its level of politeness as an answer to user's negative emotional state.

Wang et al. (2005) implemented a computational model of politeness in a tutorial dialog system. A series of Wizard-of-Oz studies showed that the polite version of the agent yielded better learning outcomes.

Glas and Pelachaud (2015b) proposed different politeness strategies for an agent according to user's engagement level.

4.4.1.3 Alignment Strategies

Kopp (2010) argued that "mechanisms like mimicry, alignment, and synchrony are essential coordination devices in face-to-face conversation" and that "it may be a significant improvement of human-agent interaction to also impart those mechanisms to embodied conversational artifacts".

A type of alignment is emotional mirroring (Acosta and Ward, 2011). Mancini et al. (2017) investigated mirroring of human laughter by an ECA during an interaction. In their experiment, a user and an agent listened to a funny music. When the ECA mirrored in real-time the behavioural expressivity of its laughter according to the user's behavioural expressivity (by taking into account trunk movements and amplitude of laughter movements), its social presence as perceived by the participants was greater than when it did not aligned its behaviour. Participants also had the feeling that it was easier to interact with the ECA, and they had the impression they were both in the same place and that they laughed together.

Campano et al. (2015b) focused on agent's verbal alignment. They built a model to select when and how the agent should generate appreciation sentences. The agent could align the lexical level of its sentences through other-repetition expressing emotional stances in appreciation sentences. Other-repetitions are the intentional repetition of a part of what the speaker just said, in order to convey a communicative function not present in the first instance. An evaluation study of the model showed a positive impact on user's engagement and agent's believability.

4.4.1.4 Reinforcement Learning

Other studies focused on how to improve user's engagement by adapting social agents (mainly robots) behaviours, using reinforcement learning (RL) methods. These works incorporated user's social signals to measure user's engagement and exploited it as the reward of the RL algorithm. For example, Ritschel et al. (2017) computed user's engagement as a reward, with the goal to adapt robot's personality expressed by linguistic style.

Gordon et al. (2016) exploited facial expressions to measure child's engagement in order to adapt a robot's behaviours, while Liu et al. (2008) exploited user's physiological signals.

4.4.2 Methods to Detect User's Engagement

In their overview, Clavel et al. (2016) distinguished three main class of methods for user's engagement detection.

Subjective methods include users' self-report of their own engagement. This can be measured through questionnaires or structured interviews. Some available questionnaires are the Engagement scale in Temple Presence Inventory of Lombard et al. (2009) or the Post-Lecture Engagement Questionnaire created by D'Mello et al. (2012). Questionnaires are easy to analyse since user's answers are often quantitative ratings on a Likert scale, but they cannot measure user's engagement within, i.e., during the interaction, since they should be intrusive. They are also prone to errors due to participants' fatigue in the case of multiple questionnaires, or due to memory problems if they are filled in after watching several stimuli. The second subjective method to measure engagement concerns the structured interviews, which allow collecting more information but are harder to analyse compared to questionnaires since they collect qualitative data. Interviews were used to evaluate Ada and Grace system described in Section 4.3 (Traum et al., 2012): these included open-ended questions and rating scale questions about users' interest, attitudes, awareness, and knowledge of themes discussed during the interaction.

Objective methods concern the detection of verbal, non-verbal and physiological signals from the user during the interaction. They are not intrusive and can measure the evolution of user's engagement during the interaction, but are prone to possible detection errors due to automatic tools. Novielli (2010) and Campano et al. (2015a) focused on speech and dialogue patterns in human-agent interaction to detect engagement through, for example, speaking turn duration and speech analysis. Ritschel et al. (2017) estimated user's engagement based on a Dynamic Bayesian Network considering head tilt and orientation, arms posture (opened or closed/crossed) and body leaning as non-verbal cues. Salam et al. (2017) exploited machine learning techniques to detect engagement in pairs or small groups of people by extracting trunk features, e.g., head/trunk orientation, quantity of movement, trunk distance. Choi et al. (2012) measured heart rate and electrodermal activity of participants when engaged in a decision task (prisoner's dilemma game) with an emotionally expressive agent. They found that electrodermal activity predicted the extent to whether people will engage affectively or strategically with the agent. Objective methods for engagement detection also include other measures, like the total time of the interaction (Bickmore et al., 2013) or the willingness to meet again the agent (Cafaro et al., 2016).

Annotated engagement is another method which consists in asking external observer to annotate user's engagement during an interaction, according to objective and subjective measures, as in Nakano and Ishii (2010); Sidner et al. (2005).

In Chapter 8 we will describe the engagement detection model that was implemented in our system to detect user's engagement in real-time. Similarly to the above works, we analyzed user's responses to the agent to determine the agent's next utterance and we extracted head/trunk non-verbal signals to compute user's attention level. Differently from them, we exploited other modalities, such as facial Action Units (see subsection 8.4.1) at the same time to compute user's engagement.

4.5 Conclusion

IN this paper we reviewed the most relevant works which addressed the topics we are interested in this Thesis. While several studies investigated the role of non-verbal behaviour in impression formation, only few of them explicitly focused on the impact of behaviours on warmth and competence dimensions. Studies investigating the effect of agent's appearance on users' impressions show that managing appearance and voice is not enough to elicit consistent impressions of agent's warmth and competence and encourage us to take into account the role of non-verbal behaviours. When looking at the studies about warmth and competence in virtual agents, their findings are in line with the phenomena we described in Chapter 3. It seems that the same patterns occur when people judge virtual agents. In particular, support for *halo effect* was found by Niewiadomski et al. (2010) and Nguyen et al. (2015), a *primacy of warmth* judgements was found by Niewiadomski et al. (2010) and results of Bergmann et al. (2012) reflect the presence of an *asymmetrical diagnosticity* of warmth and competence perception. With respect to these studies we still aim at investigating the nature of warmth and competence impressions in human-agent interaction, with a focus on the relations between the two dimensions. With respect to Bergmann et al. (2012), we considered more behaviours than only co-speech gestures, such as facial expressions, trunk leaning and head poses, and we implemented a model of impression management. To find what these behaviours are, we proposed a methodology that used natural interactions videos instead of videos of actors (see Chapter 5) with both discrete and continuous annotations.

We then reviewed the main important field studies conducted in public spaces like Science Museums. None of them dealt with impression management nor proposed a system integrating the assessment of user impressions, particularly detecting users' multimodal cues for implementing the agent's behaviour adaptation mechanisms. Moreover, they mainly focus on agent's dialogue, why in this Thesis we are interested in managing agent's non-verbal behaviour in real-time.

In one use case of this Thesis we took into account user's engagement during the interaction. We reviewed several works which focused on fostering user's engagement by using different strategies and different user's detection methods. The method that we used to detect user's engagement, described in Chapter 8, was similar to some of them but took into account both facial expressions and head and trunk rotation of the user.

The key points of this Chapter:

- It is possible to influence user's impressions of a virtual agent by managing its non-verbal behaviour.
- People tend to judge virtual agents by applying the same patterns that occur when judging humans.
- Field studies showed that user can be engaged in an interaction with a virtual agent, but they did not focus on managing agent's impressions.
- Engagement plays an important role in human-agent interaction, and can be detected and modelled by several techniques.

Positioning:

- In this Thesis, we investigated the role of non-verbal behaviour on user's impressions of agent's warmth and competence, by analysing natural human-human interaction, and considering many non-verbal behaviours than only co-speech gestures.
- We conducted a field study to evaluate a computational model for agent's impressions management based on non-verbal behaviour.
- We detected users' engagement by analysing their multi-modal behaviour in real-time.

Part IV

Warmth and Competence in Human-Human Interaction

Chapter 5

Impressions in Human-Human Interaction: analysis of NoXi database

Contents

In this Chapter we present the analysis of a corpus of dyadic natural human-human interactions between an expert and a novice. The analysis aims at investigating the relationship between observed non-verbal cues and first impressions formation of warmth and competence. We first obtained both discrete and continuous annotations of our data. Discrete descriptors included non-verbal cues such as type of gestures, arms rest poses, head movements and smiles. Continuous descriptors concerned annotators' judgments of the expert's perceived warmth and competence during the observed interaction with the novice, and they were converted into discrete variables. Then we computed Odds Ratios between those descriptors. Results highlighted the role of smiling

in warmth and competence impressions. Smiling was associated with increased levels of warmth and decreasing competence. Smiling behaviour also affected the impact of others non-verbal cues (e.g. self-adaptors gestures) on warmth and competence. Moreover, our findings provided interesting insights about the role of rest poses, that were associated with decreased levels of warmth and competence impressions.

5.1 Introduction

The first step of the approach followed in this Thesis consisted in investigating W&C perception by analyzing natural human-human interactions. The goal was to identify the non-verbal behaviours that can elicit different degrees of W&C, since in literature we found relatively few information about them (see Section 3.5). We focused on the role of type of gestures, arms rest poses (i.e., the position of the arms when not performing any gesture), head movements and smiling behaviour.

For the analysis of human-human interaction we searched for a corpus to analyse whose set up was similar to a typical interaction between a human and a virtual agent. In particular, we searched for a corpus fitting 4 criteria: we would like to analyse dyadic interactions, where participants behaved in a natural and spontaneous way, where some knowledge was shared between the participants, and where recordings of full body behaviour were available.

The results of the corpus analysis gave use more insights about W&C in human-human interaction and served as an initial set of behaviours for the virtual agent. The following step would focus on whether these signals were perceived in the same way when displayed by a virtual character.

This Chapter is organised as follows: the corpus which we analysed is introduced in Section 5.2; the methodology followed to annotate the corpus is detailed in Section 5.3; the analyses performed on the annotations and their results are described in Section 5.4 and discussed in Section 5.5.

5.2 NoXi Database

The NOvice eXpert Interaction (NoXi) database is a corpus of dyadic screen-mediated face-to-face interactions that is publicly available at https://noxi.aria-agent.eu/. It was created in the context of the H2020 project ARIA-VALUSPA (Artificial Retrieval of Information Assistants – Virtual Agents with Linguistic Understanding, Social skills and Personalized Aspects) (Valstar et al., 2016).

In each session of the database an expert and a novice discussed about a topic. The expert participant was presumed to be knowledgeable about the topic, while the novice was interested about it and wanted to learn more about it.

The 2 participants interacted from 2 separate rooms, and communicated through a big screen and a headset; a Kinect 2 was placed on the top of each screen, like shown in Figure 5.1.

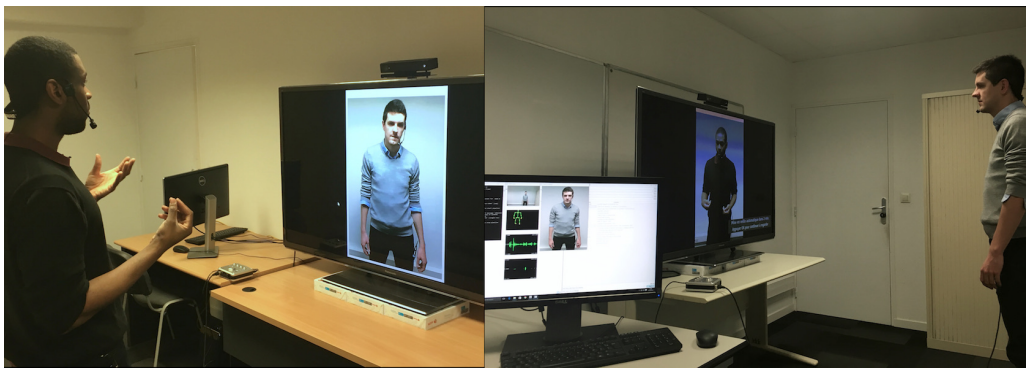


Figure 5.1 – An example of a novice-expert dyad in a recording session of NoXi database (Cafaro et al., 2017).

The choice of this specific database among others was due to several reasons. First, we considered dyadic rather than multi-party interactions available in corpora such as the Belfast storytelling dataset (McKeown et al., 2015) or the Multimodal Multiperson corpus of Laughter in Interaction (MMLI, Niewiadomski et al. (2013)), since interaction between 2 interlocutors is more similar to human-agent interaction.

Unlike other corpora, like SEMAINE (McKeown et al., 2012), where one participant adopted different roles to evoke emotional reactions, or the Interactive Emotional Dyadic Motion CAPture dataset (IEMOCAP, Busso et al. (2008)), where actors performed selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions, NoXi corpus focuses on spontaneous interactions. This was an important criterion for our study.

An available corpus focusing on natural conversation and free topic is the Cardiff Conversation Database (CCDb, Aubrey et al. (2013)), but it only recorded facial videos. NoXi database was more suitable for our purposes since its videos captured full body movements, and also the participants could see the full body of each other in the screen. This screen-mediated face-to-face interaction setup was closer to a scenario where a virtual agent is displayed on a screen.

Finally, NoXi focuses on knowledge transfer, information retrieval and occurrences of unexpected events (e.g. interruptions). We were interested in the first two aspects, too, since information sharing is one of the typical tasks performed by virtual agents. The presence of interruptions during the interactions did not affect our work since we analysed the first 5 minutes of each video (for the reasons described in the next paragraph), while interruptions were artificially injected during the recordings at about 5 minutes after it began.

	dyadic interactions	spontaneous interactions	full body recordings	knowledge sharing
Belfast Storytelling	–	√	√	–
MMLI	–	√	√	–
SEMAINE	√	–	–	–
IEMOCAP	√	–	–	–
CCDb	√	√	–	–
NoXi	√	√	√	√

Table 5.1 – A summary of the characteristics of some of the existing databases of human-human interaction, according to 4 criteria that we would need for our analyses.

5.3 Methodology

We considered the French version of the database and analyzed the videos of the "expert", since this role was more related to competence expressions, and experts were those who talked more during the dyadic interaction. We considered the first 5 minutes of the interaction for several reasons. First, we would like to prevent the participants of the videos to get used to the interaction. Second, as mentioned in the above paragraph, this choice allowed us to avoid the presence of interruptions, that were induced in the videos after the first 5 minutes. Moreover, as a first step, we decided to study the perception of W&C from non-verbal behaviour by excluding speech content. Therefore we focused only on the visual modality, leaving aside speech content and prosody features.

Two kinds of annotations were done: continuous about the perceived degree of W&C of the "expert", and discrete about non-verbal behaviours such as gestures, rest poses, head movements and smiling.

All the annotations were performed through the (Non)Verbal Annotator (NOVA) tool (Baur et al., 2015), which supports both discrete and continuous annotations (see an example of the interface in Figure 5.2). Audio of the videos was switched off when annotating as explained above.

For each annotator, we discarded the first annotated video in order to prevent any bias due to the lack of experience of the annotators with the annotation tool. In total, 14 videos, for a total of 70 minutes, were annotated. The participants of the videos were 10 men and 4 women, the 43% of them in the age range 21-25. The topic covered by the participants included history of Japan, wild animals, video games, movies, basketball, TV series, music, travels, food, South America.

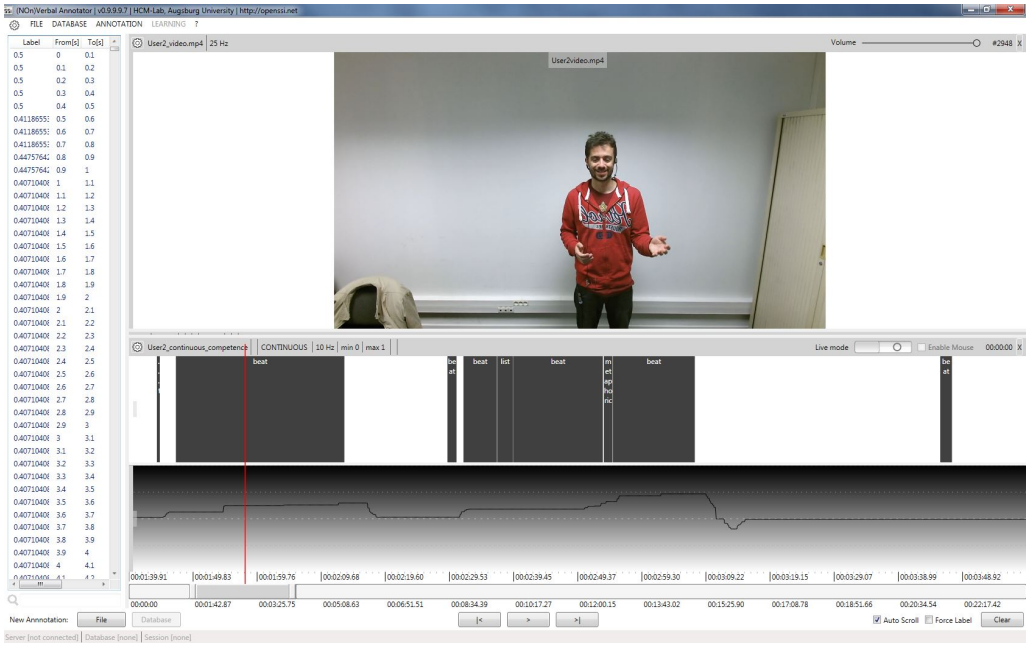


Figure 5.2 – A screen-shot of the interface for annotations in NOVA. On the top, the annotated video, in the middle a discrete annotation track and at the bottom the continuous annotation with the time line.

5.3.1 Continuous Annotations

Continuous annotations were provided by two annotators about their perceived degree of the W&C expressed by the "expert". Each dimension was separately annotated in a different time.

NOVA allowed the annotators performing live continuous annotations while watching the videos. Scores ranged from 0 (very low degree of perceived warmth or competence) to 1 (very high degree of warmth or competence), at a sampling of 25 scores per second. The continuous annotation mode was similar to GTrace (Cowie and McKeown, 2010). A white button was displayed at the left border of the track and only the value at the current playback position followed vertical mouse movement (horizontal position of the mouse was ignored). The task was easy and not tiring since it did not require to hold down the right mouse button.

For warmth annotations, the annotators were asked to evaluate how the speaker seemed "*kind, pleasant, friendly, warm towards his interlocutor*". For competence, it is to be remembered that its meaning varies according to the context of application (see Section 3.1.2). Le Deist and Winterton (2005) distinguished between functional, cognitive and social components of competence. Functional competence was not appropriate for our context, because in the database the "expert" was not performing a practical task. For the remaining two types, we chose cognitive competence, because if the expert was judged on his/her "expertise" about a topic s/he was talking about, this type of informa-

tion could be useful when modeling virtual agent's behaviour in a context of information sharing. Moreover, social competence was related to sociability, one of the main traits representing warmth. By using cognitive competence we could clearly distinguish the two dimensions (W&C) and prevent the annotation from misunderstandings. Thus, the annotator was asked to evaluate how much the speaker seemed "*knowledgeable and expert about the topic he's talking about*".

The difficulty of obtaining consistent annotations of affective content is a well-known challenge. In summary, following Metallinou and Narayanan (2013), we adopted some counter-measures: (1) the annotators were motivated and experienced people, with previous experience in affective annotation and background on literature about W&C; (2) since in literature about social cognition W&C are usually described by using a list of traits, (see Section 3.1) instead of providing a unique definition, we adopted the same approach when giving instructions to the annotators about their task, in order to make the task as clear as possible; (3) we discarded the first annotated video in order to prevent any bias due to the lack of experience of the annotators with the annotation tool; (4) we took into account the reaction lag (see subsection 5.4.1); (5) we considered the relative agreement between the annotators (see subsection 5.4.1).

5.3.2 Discrete Annotations

Discrete annotations were done at two different times, at a distance of few months, by a single annotator. A high level of agreement between the two sessions was found (Cohen's Kappa > 0.6 for each video, 29% of which > 0.8, indicating almost perfect agreement). The discrete annotations, described in the following sections, concerned: types of gestures, arms rest positions, smiling and head movements. We defined an annotation scheme for type of gestures, smiling and head movements, by being inspired from existing taxonomies and definitions (see next paragraphs). Concerning arms rest poses we created a new annotation scheme. This consisted in annotating all the rest poses present in the videos, and then kept those occurring in at least 2 videos.

5.3.2.1 Types of gestures

To define our annotation scheme for type of gestures, we combined the taxonomies proposed by McNeill (1992) and Bonaiuto et al. (2002), and we categorized gestures in 3 main groups. Table 5.2 summarizes our classification.

Beat (linked to rhythmic of the speech) and ideational (linked to the semantic of the speech) gestures are highly related to verbal expression, thus they are made only when speaking. The difference between the two categories is that ideationals are related to the semantic content of the speech, while beats are less directly so. In addition, ideationals are non-repetitive, more complex and variable in shape than beats, and they often have greater amplitude.

Label	Description
beats	Simple, repetitive, rhythmic movements that bear no obvious relation to the semantic content of the accompanying speech.
ideationals	Non-repetitive complex gestures related to the semantic content of the speech.
adaptors	Manipulations either of the person or of some object gestures; often they may serve as the basis for dispositional inferences (e.g., that the speaker is nervous, uncomfortable).

Table 5.2 – The gestures categories used in our discrete annotations and their definitions.

According to McNeill (1992), ideationals include:

- *Iconics*: they display, in their form and manner of execution, concrete aspects of the same scene that speech is also presenting. They draw their communicative strength from being perceptually similar to the phenomenon that is being talked about.
- *Metaphorics*: they are similar to iconic gestures in that they make reference to a visual image. However, the images to which they refer pertain to abstractions.
- *Deictics*: they point to a location in the gesture space.

These three subcategories are not easy to distinguish when annotating without audio, since they depend to speech content, pitch and prosody. For this reason we merged them in the ideationals category during our analyses.

Adaptors are not connected to the speech, thus they can occur at any time of the conversation and can be made by both while listening and speaking. Examples of different types of gestures are showed in Figure 5.3.

5.3.2.2 Arms rest poses

We can infer important information about others also when there are not performing gestures. Rest position and posture have been found to be possible indicators of communicator's status and attitude (Mehrabian, 1969). When "expert" did not perform any gesture (both while speaking and listening to his interlocutor), his rest poses were annotated. We

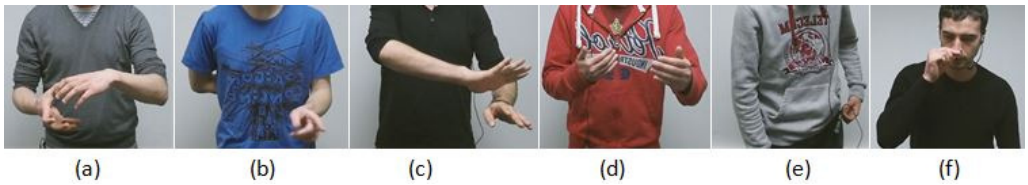


Figure 5.3 – Examples of gestures types: (a) iconic, (b) deictic, (c) metaphoric, (d) beat, (e) object-adaptor, (f) self-adaptor.

focused on arms' position during the rest pose. All the poses occurring in at least 2 videos are listed in Table 5.3 and an example of each pose is showed in Figure 5.4.

Figure 5.4 – Examples of rest poses: (a) arms_behind, (b) arms_down, (c) arms_crossed, (d) hand_inpocket, (e) hand_onhip, (f) hands_crosseddown, (g) hands_crossedmiddle, (h) hands_onhips.

5.3.2.3 Other Annotations

Head Movements. We annotated *nods*, vertical up-and-down movements of the head rhythmically raised and lowered, *shakes*, rotations of the head horizontally from side-to-side (Kapoor and Picard, 2001), and *tilts* when the expert's head tilted aside.

Smiling Behaviour. We annotated when the expert was smiling and when he was not smiling.

5.4 Data Analysis

The goal of our analyses was to investigate the presence of associations among warmth (or competence) annotations and non-verbal behaviours. As explained in subsection 5.3.2, discrete annotations of non-verbal behaviours were done by one annotator at two different times and a high level of agreement between the two sessions was found. Concerning continuous data about W&C annotations, these were done by two annotators. Before conducting the analyses we pre-processed and converted this data into discrete variables. In this way we took into account the relative agreement between the annotators and kept only the annotations were both the annotators agreed.

Label	Description
arms_behind	arms are behind the back
arms_down	arms are stretched down along the body
arms_crossed	one arm is put over the other in front of the body, so that each hand is on the opposite elbow
hand_inpocket	one hand is put into the pocket of the trousers, the other one not performing any gesture
hand_onhip	one hand on the corresponding hip, the other one not performing any gesture
hands_crosseddown	arms are laying down, hands are crossed at lower-center level
hands_crossedmiddle	similar to arms_crossed, but only hands are crossed, at center-center level
hands_onhips	two hands on the corresponding hips

Table 5.3 – The rest poses used in our discrete annotations and their descriptions.

5.4.1 Pre-processing

The full pre-processing pipeline from raw continuous annotated data to final samples used for the analysis is depicted in Figure 5.5.

One of the main issues of continuous annotations is *reaction lag*. In our context, this was the delay between the moment the impression was formed by the annotator and the motor process leading to the concrete annotation made with the mouse using NOVA (Mariooryad and Busso, 2015). We addressed this issue by shifting back 2 seconds the annotations, as recommended by Mariooryad and Busso (2015).

The second step of pre-processing was a data smoothing using a simple moving average technique, in order to reduce meaningless noise.

Figure 5.5 – Pipeline of pre-processing of continuous annotations. Data were first processed separately, then the last steps were performed only on time windows were the two annotators agreed.

When computing for agreement between two or more raters, it is recommended to consider it in relative terms rather than in absolute terms, because of each person's internal scale when assessing affective content (Metallinou and Narayanan, 2013; Yang and Chen, 2011). Therefore, before comparing the annotations of the two raters, we discretized the continuous annotations by following the approach proposed by Cowie and McKeown (2010) and applied by Chollet et al. (2014), considering the relative agreement of warmth (or competence) variations: constant, increase and decrease. Each constant was converted in the type of variation immediately preceding it, so that each variation ended when the opposite variation started. In this way, continuous annotations were converted into binary data. Figure 5.6 shows an example of this discretization.

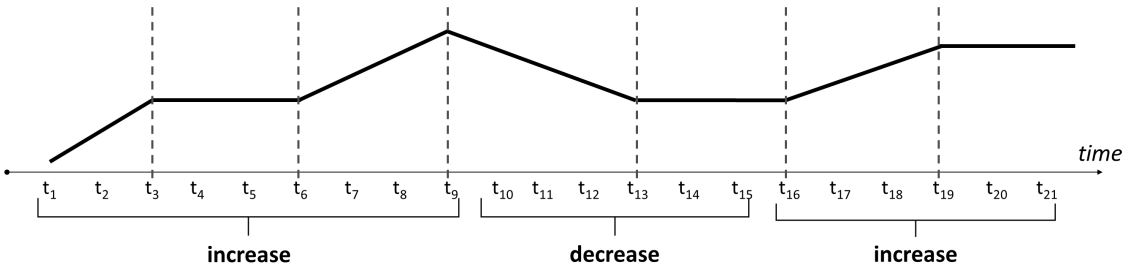


Figure 5.6 – Example of competence variation showing sampled binary discretized levels (increase vs. decrease). When constant, the sample's label for the variation was converted to the same as the one immediately preceding it.

The last steps of pre-processing were applied on discrete annotations. First, we merged the annotations coming from the two raters by keeping only the time windows where the two annotators agreed on the type of warmth (or competence) variation expressed by the expert.

Since annotations were sampled at 25 times per second, identical discrete annotations were repeated during the time windows where a non-verbal cue was performed. Thus, we shrank consecutive duplicated samples, yielding the same information, in order to avoid dependency between samples. That is, we kept only samples with at least one different

feature or placed in different time windows. The final preprocessed dataset consisted of 1087 samples for warmth and 1069 for competence.

5.4.2 Analysis and Results

In order to investigate the presence of associations among warmth (or competence) annotations and non-verbal cues, we computed Odds Ratios (ORs) (Scotia, 2010). Odds Ratios are an association measure that represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure. In our case, they represented the odds that an increase (or a decrease) of warmth (or competence) would occur given a stimulus (type of gesture, or type of rest pose, or smile, or type of head movement), compared to the odds of the decrease (or increase) of the same dimension occurring in the absence of that stimulus. That is:

$$OR = \frac{odds_{increase}}{odds_{decrease}}$$

where

$$odds_c = \frac{p_c}{1 - p_c}$$

and p_c = probability of increase ($1-p_c$ = probability of decrease) in presence of a non-verbal cue $c\epsilon$ {beat, ideational, adaptor, arms_down, arms_behind, ... , hands_crossed_middle, smile, nod, shake, tilt}.

When $OR = 1$, the presence of the stimulus does not affect odds of increase (no association between the stimulus and warmth -or competence). When $OR > 1$, the presence of the stimulus is associated with higher odds of increase (positive association). When $OR < 1$, the presence of the stimulus is associated with lower odds of increase (negative association with increase, that is, positive association with decrease). A summary of all the computed Odds Ratios is shown in Tables 5.4, 5.5 and 5.6.

5.4.2.1 Warmth

Arms Rest Poses. In general, the presence of rest poses was associated with decrease of warmth (OR of warmth increased for no_restpose vs all other rest poses $= 2.22, p < 0.01$). Indeed, when analysing each rest pose separately, a negative association with warmth was found, in a decrease order of magnitude, for arms_crossed, arms_behind and arms_down.

ORs for hands_on_hips and hands_crossed_down tended towards a positive association with warmth but they did not reach statistical significance.

Type of Gestures. Coherently with the results found for arms rest poses, an increase of warmth was more likely to be elicited in presence of gestures than in their absence. When analysing each gesture category separately, a high positive association with warmth was found for ideationals and, with smaller magnitude, for beats. No relevant association was found for adaptors.

	no_rest_poses	**arms_down**	**arms_behind**
Warmth	2.2 ****	0.60 **	0.18 ****
Competence	1.6 ****	0.80 n.s.	0.83 n.s.
	arms_crossed	**hands_crossed_down**	**hands_crossed_middle**
Warmth	0.08 **	1.36 n.s.	1.00 n.s.
Competence	0.27 ***	1.46 n.s.	1.00 n.s.
	hands_on_hips	**hand_on_hip**	**hand_in_pocket**
Warmth	3.6 n.s.	0.60 n.s.	1.23 n.s.
Competence	1.5 n.s.	0.90 n.s.	0.4 **

Table 5.4 – Odds Ratios for arm rest poses, with the correspondent p-value. (n.s. stands for p > 0.05, * for p $\leq$ 0.05, ** for p $\leq$ 0.01, *** for p $\leq$ 0.001 **** for p $\leq$ 0.0001.)

	beat	**ideational**	**adaptor**
W	1.4 *	3.09 ***	0.84 n.s.
C	1.6 **	1.3 n.s.	0.88 n.s.

Table 5.5 – Odds Ratios for types of gestures, with the correspondent p-value. (n.s. stands for p > 0.05, * for p $\leq$ 0.05, ** for p $\leq$ 0.01, *** for p $\leq$ 0.001 **** for p $\leq$ 0.0001.)

Head Movements. All head movements showed a tendency to be positively associated with warmth, but none of them reached statistical significance.

Smiling. The highest association with warmth variation concerned smiling: presence of smiling was around 9.7 times more likely to elicit warmth increase compared to absence of smiling ($p < 0.0001$).

Interaction of smiling and type of gestures. Interesting results emerged from the analysis of gestures performed with a smile or without it. In particular, for ideationals, beats and adaptors, warmth increase was more likely to be elicited when those gestures were made with a smile, compared to without smiling (all ORs > 2.4). The largest effect of smiling was for association between adaptors and warmth: when expert made an adaptor with a smile, raters always annotated warmth increase, while this occurred only in 50% of cases when adaptors were performed without smile.

Interaction of smiling and rest poses. Smiling positively affected the association of hands_crossed_middle and arms_down (ORs > 10). A warmth increase was more likely to be elicited by these rest poses when they were performed with smile, compared to those without smiling.

	nod	shake	tilt
W	1.34 n.s.	1.44 n.s.	1.74 n.s.
C	1.5 n.s.	0.94 n.s.	1.08 n.s.

Table 5.6 – Odds Ratios for type of head movements, with the correspondent p-value. (n.s. stands for p > 0.05, * for p $\leq$ 0.05, ** for p $\leq$ 0.01, *** for p $\leq$ 0.001 **** for p $\leq$ 0.0001.)

Interaction of smiling and head movements. Tilts, nods and shakes were mostly exhibited without smiling. The fewer cases when they were made with a smile were all associated with warmth increase, while in the other cases only around 50% were positively associated with warmth.

5.4.2.2 Competence

Arms Rest Poses. In general, the presence of rest poses was associated with decrease of competence (OR of competence decrease for all rest poses vs no rest poses$= 1.6, p < 0.001$). Specifically, a negative association with competence was found, in a decrease order of magnitude, for arms_crossed and hand_in_pocket.

Type of Gestures. Coherently with the results found for arms rest poses, an increase of competence was more likely to be elicited in presence of gestures than in their absence. When analysing each gesture category separately, a moderate positive association with competence was found for beats, and a moderate but no statistically significant association for ideationals. No relevant association was found for adaptors.

Head Movements. Nods showed a tendency to be positively associated with competence, but none of head movements reached statistical significance.

Smiling. Smiling was negatively associated with competence: presence of smiles was 1.6 times more likely to elicit competence decrease compared to absence of smiling ($p < 0.0001$).

Interaction of smiling and type of gestures. Smiling positively affected the association between adaptors and competence: making an adaptor with a smile was 2.21 times more likely to elicit competence increase compared to making an adaptor without smiling. Regarding other gestures, smiling had a moderate negative effect on their association with competence (OR for ideationals = 0.48, OR for beats = 0.37).

Interaction of smiling and rest poses. Smiling positively affected the association of arms_crossed with competence: competence increase was 1.6 times more likely to be elicited by this rest pose when it was performed with smile, than without smiling. In general, for the majority of the other rest poses, smiling had a negative effect on their association with competence.

Interaction of smiling and head movements. No effects were found for smiling on association between head movements and competence.

5.4.2.3 Relations between Warmth and Competence

When looking at the direction of the association of each non-verbal cue with the two dimensions of social cognition, we noted that for the majority of them an *halo effect* occurred, while an interesting *compensation effect* occurred for smiling. Results supported the *primacy effect* of warmth over competence (see Section 3.3). In most of the cases, the magnitude of the association (either positive or negative) of each non-verbal cue and warmth was higher than those of the same non-verbal cue and competence. The best evidence for a primacy effect of warmth concerned smiling. The magnitude of the association was amplified for warmth (9.67, very high) compared to competence (0.64, moderate in the opposite direction). This is in line with literature Judd et al. (2005) where magnitude of compensation effect was found to be higher for warmth compared to competence.

5.5 Discussion

Results showed the important role of smiling behaviour. Smile was associated with judgments of warmth increase and competence decrease. This is in line with previous results (Bayes, 1972; Cuddy et al., 2011), and suggests evidence of a compensation effect between the two fundamental dimensions of social cognition. Smiling also highly impacted the association of specific types of gestures and rest poses with warmth and competence judgments. For example, when experts were having their arms crossed, competence judgments decreased, but the direction of this association was reversed when the same rest pose co-occurred with a smile. We observed a similar effect between arms crossed and warmth.

The relationship between adaptors (gestures) performed while smiling with increasing competence judgments seems to be in contrast with earlier results. Self-adaptors have been often associated as displays of stress and anxiety (Ekman and Friesen, 1974), that in turn result in a low level of perceived competence. However, our result could be explained by the fact that genuine smiles softened the relationship of the self-adaptors with stress and made more prominent, for the observer, competence perception. Rest poses contributed to decrease in judgments for both dimensions. This is surprising given that in previous works there wasn't any finding linking those behaviours to W&C first impressions.

For the majority of the observed non-verbal cues (except for smiling) we found evidence in support of the halo effect. More specifically, W&C levels went towards the same direction. Results also supported the primacy of warmth over competence in terms of magnitude of effect.

As for head movements, we found some promising trends (between nods and competence's level and between tilts and warmth) but without reaching statistical significance.

5.6 Conclusion

IN this Chapter we presented the analysis of a corpus of dyadic expert-novice knowledge sharing natural interactions, for the purpose of investigating non-verbal behaviour eliciting different degrees of warmth and competence impressions. We computed the association between discrete annotations of non-verbal behaviours (type of gestures, arms rest poses, head movements, smiling) with annotations of perceived expert's warmth and competence (converted from continuous to two discrete levels describing increase and decrease).

Continuous data was pre-processed in order to take into account the reaction lag, reduce meaningless noise and consider relative agreement between the annotators rather than absolute. Only the time windows where the annotators agreed on the type of warmth (or competence) variation expressed by the expert were kept.

Results showed the important role of smiling behaviour, which was associated with judgments of warmth increase and competence decrease and highly impacted the association of specific types of gestures and rest poses with warmth and competence judgments. For the majority of the other observed non-verbal cues evidence in support of the halo effect was found, that is, warmth and competence levels went towards the same direction. Results also supported the primacy of warmth over competence in terms of magnitude of effect.

The key points of this Chapter:

- NoXi is a corpus of dyadic screen-mediated full body interactions between an expert and a novice.
- Videos of the experts were analysed by annotating non-verbal behaviours (type of gestures, arms rest poses, head movements, smiling) and perceived expert's warmth and competence.
- The presence of rest poses (gestures) was negatively (respectively positively) associated with warmth and competence.
- Beats were positively associated with warmth and competence. Ideationals were positively associated with warmth.
- Smiling behaviour was positively associated with warmth and negatively associated with competence, and highly impacted the association of specific types of gestures and rest poses with warmth and competence judgments.

Part V

Warmth and Competence Perception in Virtual Agents

Chapter 6

Warmth and Competence Perception in Videos of Virtual Agents

Contents

In this Chapter we present a perceptual study aimed at investigating how non-verbal behaviours such as the type of gestures, the frequency of gestures, the frequency of smiles and the type of arms rest poses can affect the perception of warmth and competence of a virtual agent. We also investigated the role of expectancies on

these judgements. We created videos of a virtual agent performing these different non-verbal behaviours and asked participants to rate agent's level of warmth and competence. The agent was introduced either as a puppet controlled by a human or an intelligent agent endowed with artificial intelligence. Results showed the influence of ideational gestures on warmth and competence perception, as well as the role of expectations on these effects.

6.1 Introduction

Starting from the findings obtained by the analyses described in Chapter 5, the second step of our approach was to understand whether the same processes characterizing the social perception in human-human interactions apply to ECA's perception, especially if it is possible for an ECA to express different degrees of W&C through its non-verbal behaviour.

In addition, we investigated the role of people's expectations about the agent. Burgoon (1993) stated that people have expectations about the behaviour of others during a conversation, which are primarily based on social norms and specific characteristics of the communicators. These expectations can be confirmed or violated during the interaction. Burgoon's Expectancy Violation Theory (EVT) argued that violations of these expectations generally result in more extreme outcomes compared to confirmations.

Burgoon et al. (2016) have studied the validity of their theory in the case of human-agent interaction: it seems that we have expectations about the behaviour of ECAs, and that these expectations can be violated. Their findings encouraged us to study how expectations could influence the perception of W&C of an ECA.

The study presented in this Chapter aimed to answer the following research questions:

- **(Q1a)** *Is a virtual agent perceived differently in terms of W&C according to the non-verbal behaviours it realises?*
- **(Q1b)** *If so, what are the non-verbal behaviours (or combinations of them) that allow it to be better perceived in terms of W&C?*
- **(Q2)** *Do our expectations and a-priori of an ECA influence the impressions that are formed afterwards?*

To answer these questions, we designed a perceptual study where we manipulated some non-verbal behaviours in a virtual agent. The choices regarding the signals and the design of the study resulted from the compromise between the desire to study all the signals we were interested about, and the need to limit the complexity of the design of the study.

This Chapter is organised as follows: EVT theory is described in Section 6.2; the methodology of the perceptual study, including experimental design, stimuli and procedure is detailed in Section 6.3; the analyses of participants' answers and the results are shown in Section 6.4 and discussed in Section 6.5.

6.2 Expectancy Violation Theory

This theory, proposed by Burgoon (1993), explained how humans form expectations regarding communication with other people, how they evaluate their communication experiences based on their expectations and how those evaluations of confirmed or violated expectations affect communication outcomes. Expectancies influence social interaction as they affect subsequent information processing, behaviour, and perception (Burgoon, 1993).

Expectancies are defined as cognitions about "anticipated behaviour that may be either generalized or person-specific" (Burgoon and Walther, 1990). They are a function of social norms and idiosyncrasies of the other. These last ones are based on prior knowledge of the other and reflect the extent to which the expectancies for a particular communicator deviate from the socially normative ones. With unknown others, the expectations are identical to the societal norms and standards for the particular characteristics of the communicator (e.g., her gender, age, personality), type of relationship (e.g., degree of acquaintance, status, relational history) and contextual factor of the situation. According to all these variables, people expect what behaviours are possible and appropriate.

Expectancies include an affective component, that is, they are assigned to valences. All communicators can be located on a valence continuum from positive to negative according to how "rewarding" they are seen by the observer. Target communicators expected to have positive personal qualities or to be congenial communicators have positive valence, presumably because perceivers anticipate pleasant interactions with them. Communicators attributed to be dissimilar from the perceiver have negative valence because interactions with such individuals are expected to be unpleasant. Violations of expectancies are also evaluated and they can be positive or negative. This appraisal process may be moderated by target reward valence.

Evaluation is defined as the process of assigning a valence to a violation of an expectation. Valence can be positive or negative depending on whether it is seen as favorable or undesirable. Evaluation and expectancy interact to form four EVT conditions, that are shown in Table 6.1.

The main important assumption of EVT is that positive and negative violations lead to more positive and negative interaction outcomes respectively than does conformity to expectations. Positive violations generally lead to more positive communication processes because positive violations engender greater mutuality, involvement, and interaction coordination between the violator and the target (Burgoon et al., 1999). Moreover, violations generally result in more extreme social judgments compared to confirmations (Afifi and Burgoon, 2000; Burgoon et al., 1999; Ramirez Jr and Wang, 2008).

Originally designed to explain terminal consequences of conversational distance changes during interpersonal interactions (Burgoon and Hale, 1988), EVT has been revised and ex-

Expectancy	Valence: Positive	Valence: Negative
Not expected	Positive violation (**PV**). An unexpected act, exceeding partner's expectations favorably.	Negative violation (**NV**). An undesirable, unexpected act.
Expected	Positive confirmation (**PC**). An expected and favorable act (i.e., positively valenced).	Negative confirmation (**NC**). An expected but undesirable, negatively valenced act.

Table 6.1 – The 4 possible EVT conditions, according to the combination of evaluation of valence and expectancy (Burgoon et al., 1999).

tended to apply to a greater range of non-verbal behaviours and communication outcomes (e.g., gaze).

Expectations have been found to apply also on human-computer interaction (e.g., Bonito et al. (1999)). Burgoon et al. (2016) were the first to focus on the effects of EVT conditions on social judgments in human-agent interaction. They investigated different forms of interactions, with either humans or virtual agents endowed with graduating media richness and anthropomorphism, from text-only to multi-modal animation. The task proposed to the participants was the Desert Survival Problem, which consisted in discussing with the partner (human or virtual agent) about ranking 12 items in priority to survive after being lost in the desert. Their results confirmed that humans have expectations about virtual agents and their interactions with them, and that these expectations can be violated. Thus, EVT applies to human-agent interactions. Indeed, different forms of interaction evoked varying expectations and evaluations, each reflecting one of the four EVT conditions. They used questionnaires to investigate the effect of the EVT conditions on social judgments (about the following variables: dependability, dominance, expertise, sociability, trust, and task attractiveness), communication quality and task performance. EVT predictions (PV are better than PC; NC are better than NV) applied to participants' judgments concerning some variables, like task attractiveness, but not to social judgments.

The study presented in this Chapter aimed to investigate the role of expectancies on ECAs perception. To do this, we manipulated the initial description of the agent and analysed whether this affected participants' perception of the agent's W&C.

6.3 Methodology

In order to investigate the role of non-verbal behaviours and expectancies in the perception of ECA's W&C, we conceived a perceptive study where participants were asked to

watch videos of an agent performing different non-verbal behaviours and to answer to a questionnaire about their impressions of the agent.

6.3.1 Independent Variables

The experimental design included 5 factors, one between-subject and 4 within-subjects. The choice of the non-verbal behaviours to manipulate was the result of the compromise between the desire to study the most possible signals, and the need to limit the complexity of the design of the study.

6.3.1.1 Between-subject factor

The between-subject variable concerned the *description* of the virtual agent and could take 2 values: *agent* vs *avatar*. Through this manipulation, we aimed to create 2 different expectations for the agent, similar to the work of Lucas et al. (2014) and Gratch et al. (2016). In the *avatar* condition, the virtual character was introduced as “a puppet that communicates with others, whose gestures, facial expressions, and dialogue are controlled by a human operator”. In the *agent* condition, the character was introduced as “a virtual agent endowed with artificial intelligence and able to communicate with others, whose gestures, facial expressions and dialogue are controlled by a series of algorithms ". Two different images were associated to the descriptions, as shown in Figure 6.1.

Figure 6.1 – The two images associated to the initial descriptions of the virtual agent: the first was shown in the *agent* condition, while the second one was shown in the *avatar* condition.

To verify that the participants had understood and retained the description of the virtual character, a question of control was asked at half time of the experiment, where the participant must choose the right answer between the description given at the beginning of the event experience, and the one relating to the other condition.

6.3.1.2 Within-subject Factors

The within-subject variables were:

- *Type of gestures*: gestures made by the agent could be *Beats* -rhythmic gestures unrelated to the speech content- or *Ideational* gestures -more complex gestures related to the content of the speech.

- *Frequency of gestures*: during the video, the agent performed 3 series of gestures (*HighFreqGestures*), or only one (*LowFreqGestures*).

- *Frequency of smiles*: During the animation, the agent smiled 3 times (*HighFreqSmile*) or only once (*LowFreqSmile*).

  The possible locations of smiles were at the beginning, middle and end of the video. To choose when to display the smile in the *LowFreqSmile* condition, we performed a manipulation check (N = 14, including 6 women) to verify whether the moment of appearance of the smile influenced the perception of the frequency of the smiles in the video. We showed each participant a video zoomed on the agent's face, which smiled according to the corresponding condition (3 times, 2 times or only once in all possible locations) and at the end we asked them to note how many smiles the agent displayed. Apart from the condition *HighFreqSmile*, where the high frequency of smiles was recognized, the condition with only one smile at the end was the only one between those of *LowFreqSmile* to be well recognized (all participants noted 1 smile), while in the other conditions the participants noted several smiles, and under the conditions with 2 smiles, most noted more than 2 smiles (up to 7). That's why we chose to keep the smile in the final position for the *LowFreqSmile* condition.

- **Type of arms rest poses**: when the agent did not display communicative gestures, its arms were at rest; it *Crossed* them or put its hands on the hips with its elbows bent outward (this is called *Akimbo*).

  In order to limit the complexity of the design, we chose the two arms rest poses that seemed the most interesting. *Crossed* was the only one, between those analyzed in our previous study, to be linked (negatively) to both W&C. *Akimbo* was a pose much studied in the literature in relation to the dominance expressed by a human (Ball and Breese, 2000) and also by an ECA (Straßmann et al., 2016).

  In the final version of the videos, the arms rest poses appeared only at the beginning and at the end of the animation, since the quality of the animation of the agent did not allow several switches from a gesture to a particular rest pose. During the video, when the agent did not perform any gesture, its arms remained along its body.

The different combinations of these 4 variables therefore gave 16 experimental conditions.

6.3.2 Dependent Variables

After each video, the participants were asked to rate:

- Their perception of agent's warmth, by rating on a 5-points Likert scale how much they agreed that the agent was *kind, pleasant, friendly, warm*;
- Their perception of agent's competence, by rating on 5-points Likert scale how much they agreed that the agent was *competent, effective, skilled, intelligent*.

  The adjectives came from the two scales selected by Aragonés et al. (2015), which showed that these adjectives have a high degree of reliability when used to describe the perception of individuals.

6.3.3 Hypotheses

Concerning the between-subject variable, we hypothesised **(H1)** that the different descriptions of the agent would gave 2 different expectations regarding its competence and, due to *halo effect*, warmth level, and therefore would influence their answers to the questionnaires (subsection 6.3.2). The effect of expectations could be expressed globally in the results or interact with other independent variables.

Moreover, we hypothesised that the agent would be perceived differently in terms of W&C depending on the type of gestures it realised. Following the results of our previous study, where *Ideationals* had a greater magnitude of association than *Beats*, we hypothesised **(H2a)** that when the agent used ideational gestures it would be perceived as warmer than when performing beats. We also hypothesised **(H2b)** that there would be a similar effect of these gestures on competence ratings, as shown in the literature (Maricchiolo et al., 2009).

We hypothesised that the agent would be perceived differently in terms of W&C according to the frequency of gestures. Following the results of our previous study, in which the presence of gestures was positively associated with both W&C, we hypothesised that by increasing the presence of gestures (and therefore their frequency) the agent would be perceived as warmer **(H3a)** and more competent **(H3b)** than when performing gestures with low frequency.

Our analyses presented in Chapter 5 showed the presence of a positive association between smiling behaviour and warmth and a negative association with competence. We hypothesised that the frequency of smiles would influence the perception of the agent: the more it smiled, the warmer **(H4a)** and the less competent **(H4b)** it would be perceived.

Concerning the arms rest poses, since the signals expressing competence are also related to dominance (Cuddy et al., 2008), we hypothesised that **(H5b)** the agent would be more competent when it performed *Akimbo*. Regarding warmth, we hypothesised that **(H5a)** when the agent had its arms *Crossed*, it would be perceived as less warm.

6.3.4 Stimuli

Each video represented the virtual agent giving advice about traveling. The animation was realized thanks to the platform GRETA/VIB (Pecune et al., 2014). The gestures and the dialogue were taken from an extract of a video from the NoXi database (see subsection 5.2). To limit possible effects of the dialogue, speech and prosody were always the same, while only non-verbal cues were manipulated (see Section 6.3.1.2). The duration of each video was 20 seconds. Some examples of the agent and its behaviours are shown in Figure 6.2.

Figure 6.2 – Some examples of non-verbal behaviours realised by the virtual agent: a beat gesture, an ideational gesture, arms crossed, the akimbo position, and the agent when smiling (close-up).

6.3.5 Procedure

The experiment was available online; all instructions and questionnaires were in English, as well as the agent's dialogue. Data were collected from 32 participants (including 17 women), 18 assigned to the condition *avatar* (*Group 1*) and 14 to the condition *agent* (*Group 2*). The average age was 27$\pm$3.6, with a majority of French and Italians, but with a sufficient level of English to participate. Only participants who completed the entire experience were considered in the analysis.

The total duration of the experiment was about 20 minutes. Before starting, after reading and approving the consent form, participants read the experience scenario as well as the description of the agent that was either *agent* or *avatar* depending on the group they were assigned to. Then they could read the instructions of the experiment.

During the experiment, each participant watched 16 videos corresponding to the combinations of the independent variables. The order of appearance of the videos was counterbalanced thanks to a 16$\times$16 Latin square to limit possible undesirable effects on ratings.

After each video, participants were invited to answer questions about the perceived W&C of the agent. After answering, they could move to the next video, and so on. Between the eighth and the ninth video, the verification question (see Section 6.3.1.1) was

displayed on the screen, and the participant had to answer it in order to continue the experiment.

The last part of the experiment concerned the collection of demographic information.

6.4 Analysis and Results

According to literature (Aragonés et al., 2015), and after finding very high Cronbach's alpha coefficients, we grouped the scores related to competence ($\alpha \geq$0.88 for each video, $\mu =$0.94) and those related to warmth ($\alpha \geq$0.87 for each video, $\mu =$0.92).

As data satisfied the required ANOVA's assumptions, 2 mixed $2 \times 2 \times 2 \times 2 \times 2$ ANOVAs were performed, one for each dependent variable (warmth or competence).

6.4.1 Warmth

Warmth scores were subjected to a mixed ANOVA with *description* (*agent* vs *avatar*) as the between-subject variable and the other 4 within-subject variables (see subsection 6.3.1.2).

The ANOVA revealed two statistically significant main effects: that of the *Type of gestures* ($F(1, 28) = 18.38$, $p < 0.01$) and that of the *Frequency of gestures* ($F(1, 28) = 12.52$, $p < 0.01$). The interaction between these 2 variables was significant too ($F(1, 28) = 4.81$, $p < 0.05$).

The main effect of the *Type of gesture*, shown in Figure 6.3a, indicated that the perceived warmth ratings were statistically higher for *Ideationals* ($M = 4.07$, $SD = 1.37$) than for *Beats* ($M = 3.7$, $SD = 1.29$). This result supported **H2a**. The main effect of the *Frequency of gestures* indicated that the perceived warmth ratings were statistically higher for *HighFreqGestures* ($M = 4.06$, $SD = 1.32$) than for *LowFreqGestures* ($M = 3.72$, $SD = 1.34$). This result supported **H3a**.

The interaction between *Type of gestures* and *Frequency of gestures* showed that the maximum warmth's ratings were those of videos where the agent performed *Ideational* gestures at *HighFreqGestures* and that the *Type of gestures* affected the perception of warmth only when it was performed with a high frequency. Figure 6.4b shows the interaction effect between *Type of gestures* and *Frequency of gestures*.

Another significant interaction, but quite complex, emerged from this analysis: it was the interaction between *Type of gestures*, *Frequency of gestures*, *Type of rest position* and *description*.

The results did not allow us to validate the hypotheses **H4a** and **H5a**, relative to possible effects of *Frequency of smiles* and *Type of arms rest poses* on perceived warmth.

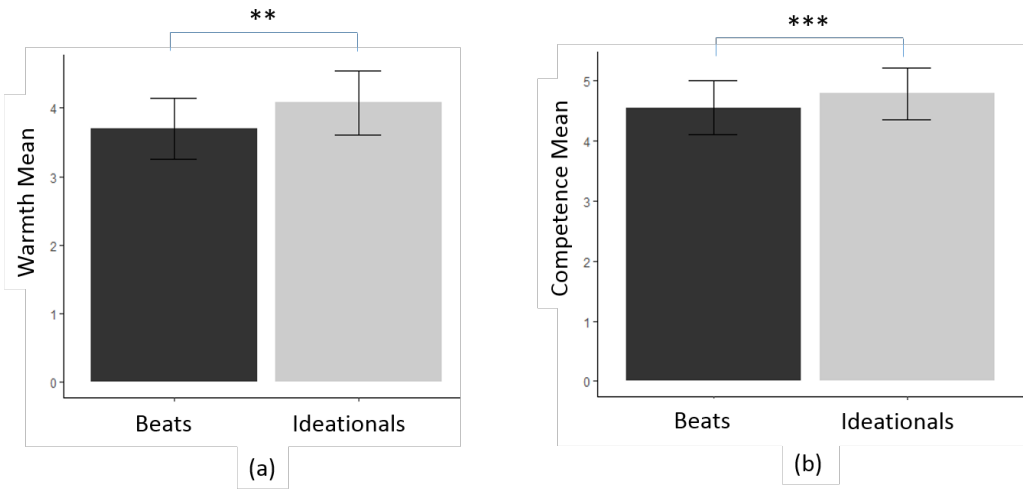


Figure 6.3 – Main effect of *Type of gestures* on (a) warmth and (b) competence ratings. N.S. stands for p > 0.05, ** for p $\leq$ 0.01, *** for p $\leq$ 0.001.

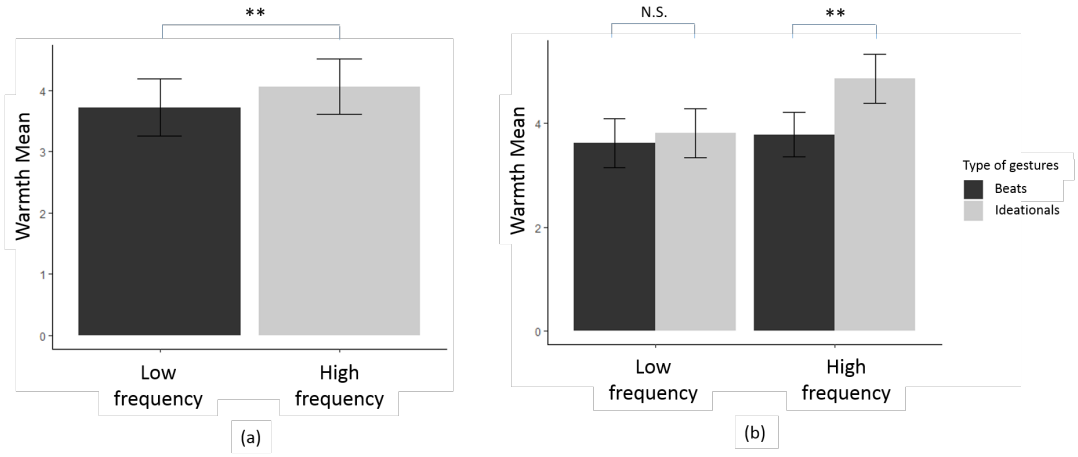


Figure 6.4 – (a) Main effect of *Frequency of gestures* on warmth ratings and (b) interaction between *Type of gesture* and *Frequency of gestures* on warmth ratings. N.S. stands for p > 0.05, ** for p $\leq$ 0.01.

6.4.2 Competence

Perceived competence ratings were submitted to a mixed ANOVA with *description* (*agent* vs *avatar*) as the between-subject factor and the other 4 within-subject factors (see subsection 6.3.1.2).

A main effect of *Type of gesture* was found ($F(1, 28) = 18.92$, $p < 0.001$). In particular, as shown in Figure 6.3b, the perceived competence ratings were statistically higher for an agent performing *Ideational* gestures ($M = 4.79$, $SD = 1.25$) than *Beats* ($M = 4.55$, $SD = 1.32$). This result supported **H2b**. No interaction effect between the factors was

significant. The results did not allow us to validate the hypotheses **H3b**, **H4b**, and **H5b** about a possible effect of *Frequency of gestures*, *Frequency of smiles* or *Type of arms rest poses* on perceived competence.

6.4.3 Effect of Agent's Description

We then analysed participants' ratings by dividing them into groups according to *description*, i.e., whether the agent was introduced as an *avatar* or an intelligent *agent* (see subsection 6.3.1.1). We found that the significant effects described in the previous paragraphs were significant only for participants in *agent* condition, while no significant effects were found for participants in *avatar* condition. In particular, the main effect of *Type of gestures* on competence scores was statistically significant in *agent* condition ($F(1, 21) = 18.11$, $p < 0.001$) and not present in *avatar* condition ($F(1, 9) = 2.34$, $p > 0.1$). Concerning warmth's ratings, the main effect of *Type of gestures* existed in *agent* condition ($F(1, 21) = 13.27$, $p < 0.01$) and not in *avatar* condition ($F(1, 9) = 0.38$, $p > 0.5$); the main effect of *Frequency of gestures* existed in *agent* condition ($F(1, 21) = 8.92$, $p < 0.01$) and not in *avatar* condition ($F(1, 9) = 3.19$, $p > 0.1$); the interaction between *Type of gestures* and *Frequency of gestures* existed in *agent* condition ($F(1, 21) = 6.78$, $p < 0.05$) and not in *avatar* condition ($F(1, 9) = 0.16$, $p > 0.7$).

These results supported hypothesis **H1**, since they showed an influence of *description* variable on participants' ratings.

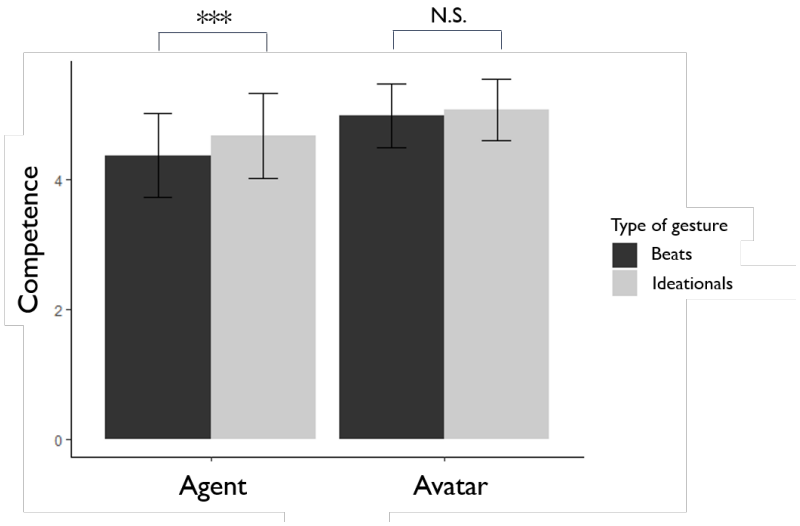


Figure 6.5 – Effect of agent's *description* on the effect of *Type of gestures* on competence ratings. *** stands for $p < 0.001$, N.S. for $p > 0.05$.

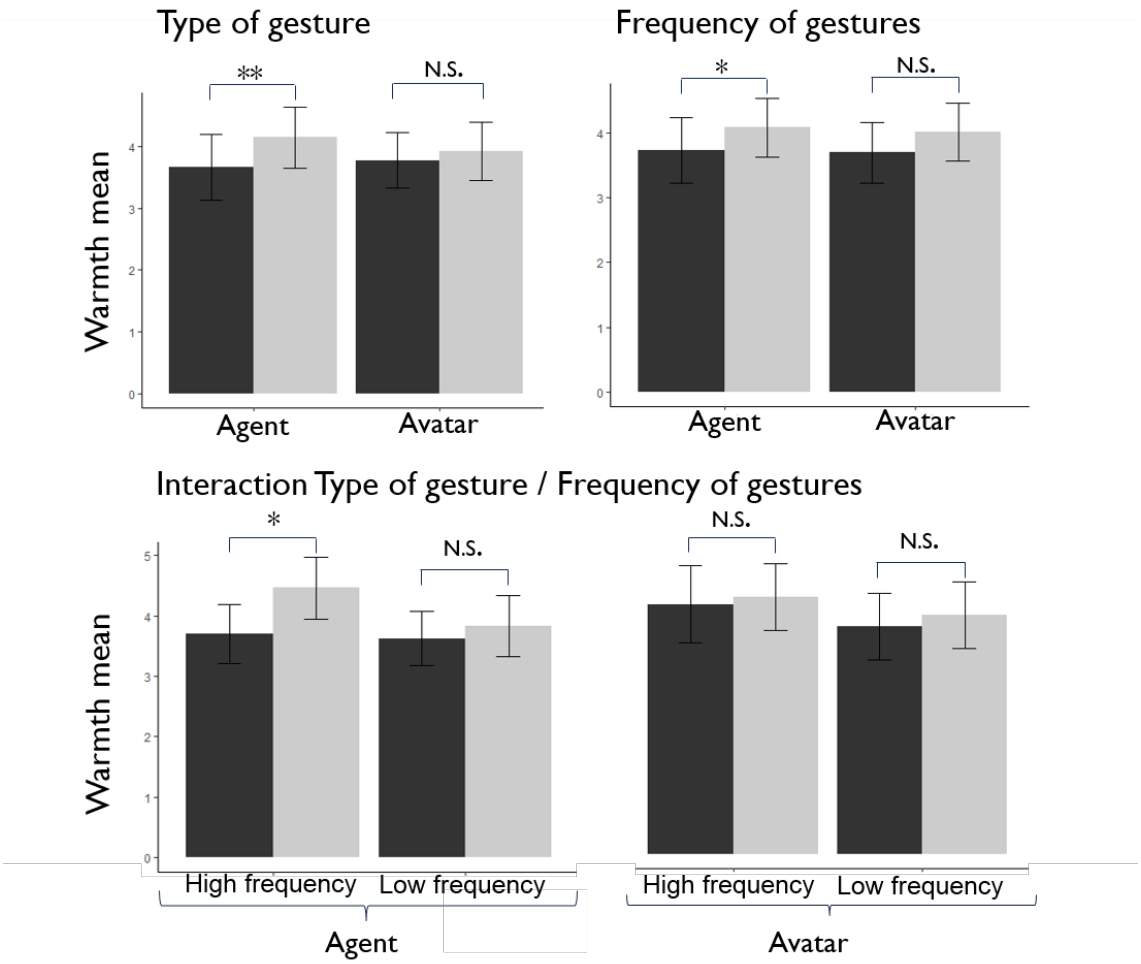


Figure 6.6 – Effect of agent's *description* on the effects of *Type of gestures*, *Frequency of gestures* and their interaction on warmth ratings. * stands for $p < 0.05$, N.S. for $p > 0.05$.

6.5 Discussion

The results of this perceptual study showed the influence of the type of gesture on both the perception of W&C, in particular when the agent made ideational gestures (related to what it was talking about) it was perceived as warmer and more competent compared to when performing beat gestures whose forms were not related to the content of its speech. The use of these gestures may reflect the motivation of the agent to help the user better understand what it was talking about, and at the same time, its knowledge of the topic.

With regard to warmth, ideational gestures had a positive effect on the perception of this dimension only when they were made at high frequency.

Concerning the hypothesis of an effect of the expectations on the judgments on the agent, when the agent was presented as intelligent and autonomous, this affected participants' ratings compared to when the agent was presented as a puppet controlled by a

human. This result seemed to support the role of people's expectancies on impression formation, and highlighted the importance to take into account participants' a-priori about virtual agents in our further studies.

No effect of the frequency of smiles was found. This could be explained by the fact that the presence of a single smile was already sufficient to give an impression and that it did not vary if one increased the frequency of this signal.

Finally, we did not find any effect of the arms rest poses. Actually, for animation reasons (see subsection 6.3.1.2), the agent performed a specific rest pose only at the beginning and at the end of the video, while during the rest of the video the rest pose consisted of the arms along its body. Probably the presence of the different rest poses under investigation may be too subtle to perceive a difference between the conditions.

It is interesting to compare these results with those obtained in the study of human-human interaction described in the previous Chapter. In particular, similarly to the previous results, we found a *halo effect* for participants' judgments about gestures, as they went in the same direction for warmth and competence direction. In addition to the previous study, here we found an effect of ideational gestures on competence perception, while previously this non-verbal behaviour did not result significantly associated with this dimension. In contrast to results of the analysis of human-human interaction we did not find any effect of smiling on user's perception of agent's warmth and competence, neither the *compensation effect* that we found in the previous study.

6.6 Conclusion

THIS Chapter described the methodology and results of a perceptual study inspired from the findings of the analysis of human-human interaction. The goal of the study was to investigate whether the non-verbal behaviours associated to different degrees of warmth and competence in human-human interaction were perceived in the same way when realised by a virtual agent. Results showed that the agent was perceived as warmer and more competent when performing ideational gestures compared to when performing beats, and that it was perceived as warmer when performing ideationals at high frequency. All these effects were found only for participants to which the agent was described as an intelligent agent, while they were not found for participants to which the agent was described as a puppet controlled by a human. No effect of frequency of smiles or type of arms rest poses was found, but this could be due to the design of the stimuli. This was in contrast with previous results found from our analysis of human-human interaction, were a compensation effect was found for smiling on warmth and competence judgments. On the other hand, similarly to previous results, a halo effect was found for gestures on warmth and competence judgements.

The key points of this Chapter:

Contributions :

- A perceptual study was run, where non-verbal behaviour and initial description of a virtual agent were manipulated.
- The agent was perceived as warmer and more competent when performing ideational gestures, compared to when performing beats.
- The agent was perceived as warmer when performing ideational gestures at high frequency.
- The initial description of the agent seemed to have contributed to form expectations about the agent and influenced participants' ratings about agent's warmth and competence.

Part VI

Warmth and Competence in Human-Agent Interaction

Chapter 7

System Architecture for Agent's Impression Management

By seeking and blundering we learn.

Johann Wolfgang von Goethe

Contents

THIS Chapter gives an overview of the system architecture of agent's impression management aiming to endow an Embodied Conversational Agent with the capability of adapting its impressions according to user's reactions. The architecture includes three main modules, which allow for the detection and interpretation of user's verbal and non-verbal behaviour, the learning about which impression to elicit, and the final animation of the agent.

7.1 Introduction

The main goal of this Thesis was to build a computational model for an Embodied Conversational Agent able to manage its impressions of W&C towards the user. In this Chapter we present the architecture we conceived to endow the ECA with the capability of adapting its behaviour to user's reactions. The architecture is general enough to allow for customisation of its different modules according to different contexts and goals of the agent. Two examples of a personalised instantiation of this architecture are presented in Chapter 8 and 9, where the architecture was customised with the purpose of adapting agent's behaviour according to user's engagement and user's impressions of the agent, respectively.

The work presented in this Chapter was realized in collaboration with the Master student Paul Lerner[1] and Professor Maurizio Mancini[2].

The main goal of the model was to manage agent's non-verbal behaviour to elicit different impressions of W&C according to user's reactions. To do this, 3 main modules were involved:

1. The *User's analysis module* for detecting and interpreting user's reactions;
2. The *Impressions management module* for selecting the impression to elicit, through agent's communicative intentions;
3. The *Agent's behaviour generation module* for the animation of the agent.

We included a Dialog Planner in the *Impressions management module*, even if we tried to keep the dialogue as basic as possible, since we were not focusing on it.

As we said previously, the goal of the agent was to adapt its behaviours to each participant. This implied endowing the agent with the capability to learn in real-time what was the best behaviour to perform, according to its goal (e.g., elicit warmth-related impression) and user's reactions (e.g., user's impression about agent's warmth).

In a context such as human-agent interaction it is often impossible to have examples of the desired behaviour representing all the situations in which the agent could act. Instead, the agent has to learn from experience. Thus we searched for a method that would allow the agent to learn in an interactive environment by trial and error, without requiring previous knowledge about the user, with the goal to maximise user's impressions about the agent.

Supervised learning was not the best method to use in such a context. Indeed, it is defined as "learning from a training set of labeled examples provided by a knowledgeable external supervisor" (Sutton and Barto, 2018). Knowledge is provided about which action corresponds to a situation. The goal of supervised learning is to generalize a set of rules

[1] UFR de Mathématiques et Informatique, Université Paris Descartes
[2] School of Computer Science and Information Technology, University College Cork

from the provided knowledge in order to make decisions for new situations not present in the training set.

Unsupervised learning was not the best approach either. Even if this class of methods deals with unlabeled data and so does not rely on examples of correct behaviour, they are used to uncover hidden structures in the data (e.g., clustering) rather than maximizing a reward. Reinforcemel Learning (RL) thus seemed the best suitable approach for our needs. RL does not require previous knowledge of the environment and has the goal of maximizing a reward instead of uncovering hidden structures in the data. The typical framing of RL includes a loop where an agent takes *actions* in an *environment*, these actions are interpreted into a *reward* and a representation of the *state*, which are fed back into the agent. This well fits our general framework where the agent would perform behaviours in the environment of the interaction states, receive a reward from users' reaction and use it to adapt its behaviour. As RL allows the agent to "learn from interaction" (Sutton and Barto, 2018), it faces the challenge of finding a balance between exploration and exploitation. That is, the agent has to exploit what it has already experienced in order to obtain reward, but it also has to explore in order to make better action selections in the future. The dilemma is that neither exploration nor exploitation can be pursued exclusively without failing at the task. The agent must try a variety of actions and progressively favor those that appear to be best.

In the next Sections we will first describe the overall architecture of our system and then give more details about its 3 main modules.

7.2 Overall Architecture

We implemented software modules to capture user's behaviour (speech, gaze, facial expressions, head and trunk orientation), analyse/interpret it (e.g., detect the user's impressions of the agent) and decide what the ECA should say and how (i.e., the non-verbal behaviours accompanying speech).

Figure 7.1 illustrates the system we designed and implemented. We can distinguish 3 main parts:

1. *User's analysis.* We exploited the EyesWeb platform (Camurri et al., 2004) to extract in real-time: (1) user's non-verbal signals (e.g., head and trunk rotation) starting from the Kinect depth camera skeleton data; (2) user's face Action Units (AUs), by running the OpenFace framework (Baltrušaitis et al., 2016); (3) user's gaze, thanks to the eye tracker Tobii; (4) user's speech, by executing the Microsoft Speech Platform[3]. These low-level signals were processed by EyesWeb and other external tools, such as machine learning pre-trained models, to extract high-level features about the user.

[3] https://www.microsoft.com/en-us/download/details.aspx?id=27225

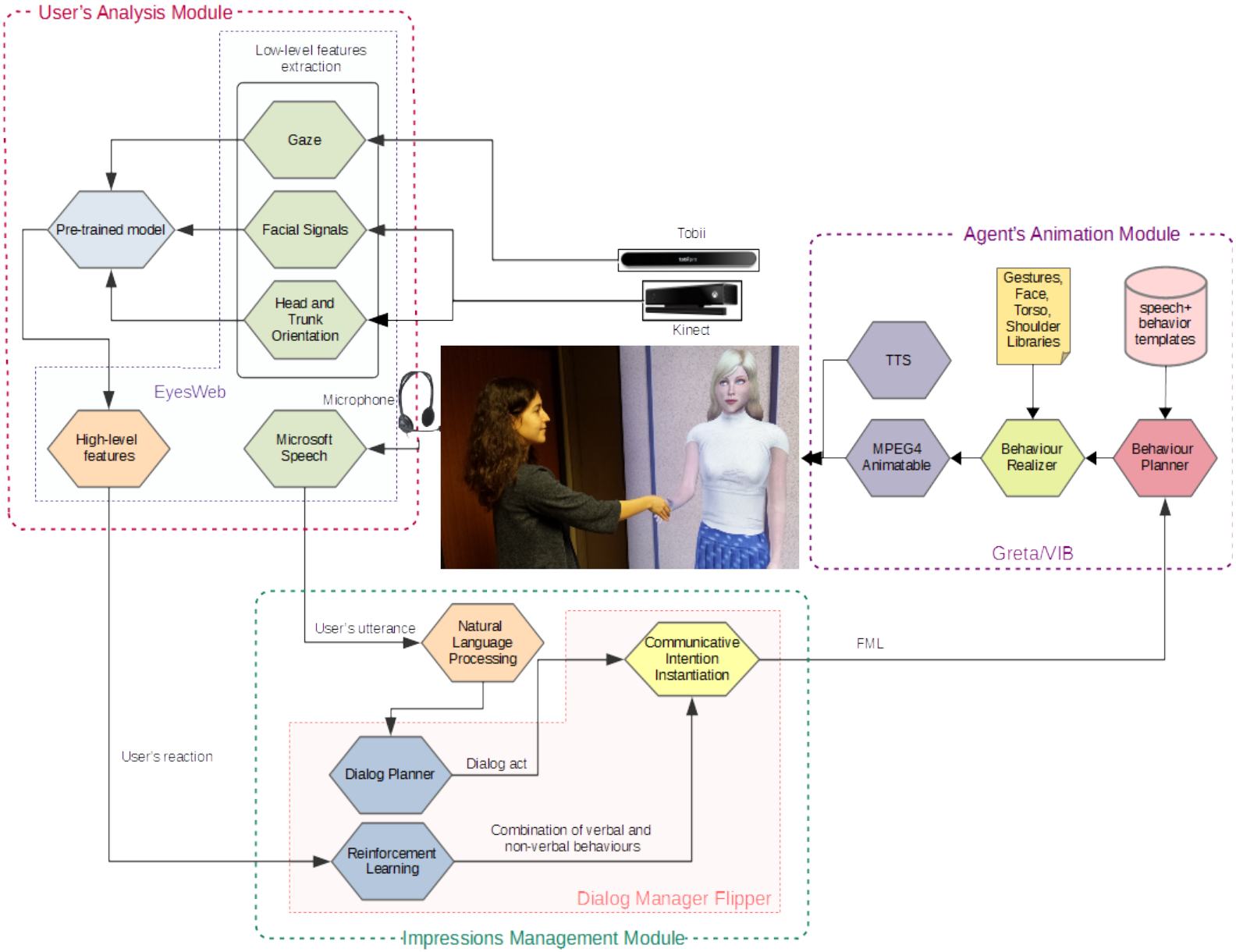


Figure 7.1 – System architecture: in *User's Analysis Module* user non-verbal and verbal signals are extracted by EyesWeb and the Microsoft Speech Platform, respectively; high-level features of the user are sent to the *Impressions Management Module* where a reinforcement learning algorithm and a dialog planner select the communicative intention that will be performed by the agent thanks to the *Agent's Generation Module*.

2. *Impressions management.* It was the decision making module of the system, where user's information was exploited by a RL algorithm and user's speech could be processed by natural language processing tools and sent to a dialog manager. The output of the module were the verbal and non-verbal behaviour eliciting different levels of W&C according to user's goal.

3. *Agent's Animation.* Agent's behaviour generation was performed by VIB/Greta, a software platform supporting the creation of socio-emotional embodied conversational agents (Pecune et al., 2014). VIB/Greta generated the ECA animation consisting of gestures, facial expressions and gaze, in synchrony with speech.

7.3 User's Analysis Module[4]

In this module we exploited EyesWeb XMI, an open software platform that supports the design and development of real-time multimodal systems and interfaces. EyesWeb is designed and developed by InfoMus Lab of University of Genova[5].

EyesWeb managed the data coming from the sensors, extracted low-level signals such as trunk and head orientation and computed mid-level features (e.g., body and head attention over time). Internally, it exploited OpenFace to extract Action Units (AUs) from Kinect's RGB information. Facial expression is one of the main non-verbal channels humans use to communicate emotions (Ekman, 2002). The Facial Action Coding System (FACS) is an annotation system for human facial actions (Ekman, 2002). Facial expressions are encoded as the composition of several Action Units (AUs) that describe the contraction of different muscles/regions of the face (e.g., inner brow raiser, cheek raiser, lip corner puller, etc.).

Figure 7.2 shows an example of the EyesWeb platform interface reporting some of the user's behaviour parameters and the detected user's AUs.

7.3.1 OpenFace

OpenFace is an open source tool intended for computer vision and machine learning researchers (Baltrušaitis et al., 2016). The software is available for download at GitHub[6]. It is capable of facial landmark detection, head pose estimation, facial action units recognition, and eye-gaze estimation. AUs detected by OpenFace are listed in Table 7.1.

In our *User's analysis module* OpenFace was exploited by EyesWeb in order to extract facial AUs that would serve as facial descriptors for high-level variables, such as user's impressions.

The tool offered two kinds of scores for the AUs (see Table 7.1):

[4]This work was realised in collaboration with Prof. Maurizio Mancini.
[5]http://www.infomus.org/EyesWeb_eng.php
[6]https://github.com/TadasBaltrusaitis/OpenFace

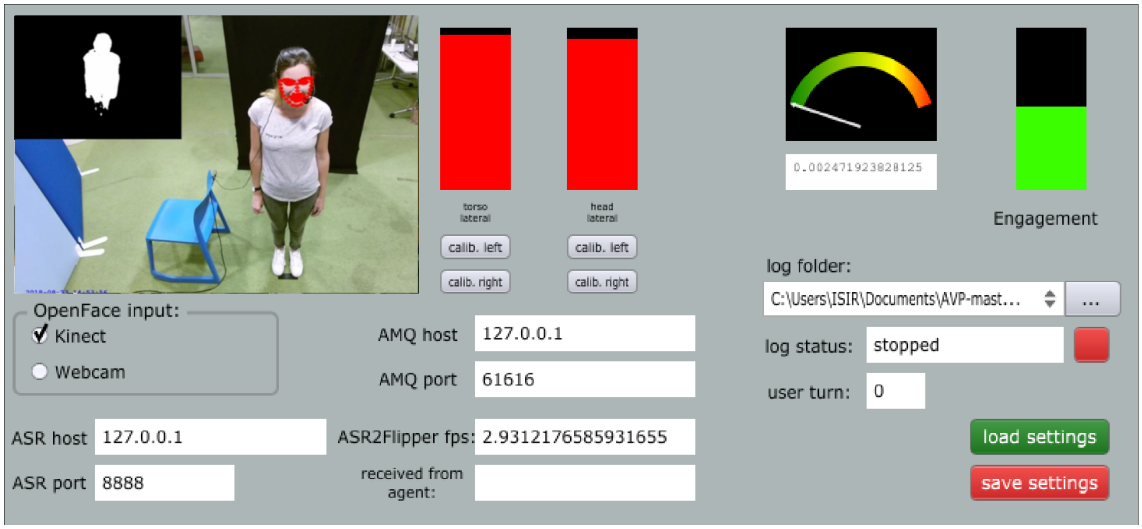


Figure 7.2 – An example of EyesWeb interface. On the left, the user's silhouette is extracted from Kinect's depth image data. The two red bars in the middle indicate that the user is looking at the screen, with both her trunk (left bar) and head (right bar). Audio intensity is very low (volume meter on the right), that is, the user is not speaking. Finally, a high-level feature, in this case user's engagement level (between 0 and 5), is represented by the green bar on the right.

AU	Description	AU	Description
1	Inner Brow Raiser	14	Dimpler
2	Outer Brow Raiser	15	Lip Corner Depressor
4	Brow Lowerer	17	Chin Raiser
5	Upper Lid Raiser	20	Lip Stretcher
6	Cheek Raiser	23	Lip Tightener
7	Lid Tightener	25	Lips Part
9	Nose Wrinkler	26	Jaw Drop
10	Upper Lip Raiser	28	Lip Suck
12	Lip Corner Puller	45	Blink

Table 7.1 – List of AUs detected using OpenFace.

1. Presence: it indicated the presence or absence of 18 AUs.
2. Intensity: the intensity of 17 AUs on a continuous value scale from 1 (minimally present) to 5 (present at maximum intensity); a score of 0 indicated absence.

The *User's analysis module* allowed the implementation of sub-modules to perform machine learning algorithms on the raw signals in order to compute high-level information about the user. In this Thesis we implemented a module to detect user's engagement from

user's AUs and head and trunk rotation (see Chapter 8) and a module to detect user's impressions from user's AUs (see Chapter 9). The output of these sub-modules was then integrated in EyesWeb, together with the other inputs from sensors.

The outputs of *User's analysis module* consisted in user's transcripted speech and user's high-level information coming from the interpretation of raw signals made by EyesWeb and the more sophisticated sub-modules. These output were sent by EyesWeb to the *Impressions management module*.

7.4 Impressions Management Module[7]

This module was implemented in the Dialog Manager Flipper, an open-source engine for pragmatic yet robust interaction management for ECAs (van Waterschoot et al., 2018).

The main components of this module were the *Dialog Planner* which selected the dialog act to perform, and the *Reinforcement Learning algorithm* that decided *how* to perform the dialog act.

The output of this module was the communicative intention that would be realised by the agent thanks to the *Agent's animation module*.

7.4.1 Flipper

The Dialog Manager Flipper is based on two main components described in XML: the *information state* and the *declarative templates*. The information state stores interaction-related information and data in a hierarchical tree-based structure. Declarative templates can be grouped and organized in different files according to their related functionality (van Waterschoot et al., 2018). Each template consists of:

- *preconditions*: sets of rules that describe when a template should be executed;
- *effects*: associated updates to the information state.

So, for example, we defined a template whose *precondition* was that if the user's impression of agent's warmth had been computed by EyesWeb and the *effect* was that the expected reward of the current communicative intention had to be updated depending on the value of this impression. Figure 7.3 shows an example of a template used in our system.

Flipper communicated with the Behaviour Planner of the ECA by sending Functional Markup Language (FML) information (see Section 7.5).

Flipper could also be exploited to implement a tool based on natural language processing (NLP), aiming at interpreting user's speech. Since the generation of a realistic and complex dialogue was not the main focus of this Thesis, we only developed a simple

[7]This work was realised in collaboration with Paul Lerner.

```
<template id="video_games_" name="video_games_positive" conditional="true">
        <preconditions>
            <condition>is.nlu.opinion=="positive"</condition>
            <condition>
                <![CDATA[helpPrint("SELECT MOVE : "+is.dialogue.step.current+is.nlu.opinion+is.states.agent.strategy)]]>
            </condition>
        </preconditions>
        <effects>
            <assign is="is.states.agent.bestMove">{ "id" : is.dialogue.step.current,"relevance" : 1}</assign>
            <assign is="is.states.agent.fml.template">is.dialogue.step.current+is.nlu.opinion+is.states.agent.strategy</assign>
            <assign is="is.dialogue.timer.silence">is.dialogue.timer.default</assign>
            <assign is="is.dialogue.step.current">"simulator_intro_1"</assign>
        </effects>
</template>
```

Figure 7.3 – An example of a template of the Dialog Manager Flipper representing the reply of the agent to a positive answer of the user to a question about video games. The precondition is that the polarity of user's speech is positive and the effect consists in three steps. First, the selection of the FML template including the dialog act and the communicative strategy that the agent has to perform; second, the threshold of user's silence that the agent has to wait before continuing to talk is set to the default length (1.5 seconds); finally, the next dialog act is selected.

tool to detect the polarity of user's answers (that is, to distinguish between negative and positive answers) rather than the semantic content of user's speech.

The *Dialog Planner* had also the task of detecting when each speaking turn started (i.e., when the agent started to speak) and ended (i.e., when the user ended to speak or when a silence threshold passed). According to the type of dialog act performed by the agent, we defined different silence thresholds. If the dialog act included a question, the threshold was longer, while if the dialog act concerned an explanation the threshold was shorter. If no voice activity was detected by the Microsoft Speech Platform during that period of time, the *Dialog Planner* started the next template. At the end of each speaking turn, the RL algorithm was run and the next communicative intention was selected.

7.4.2 Reinforcement Learning

We implemented in Flipper a decision making component, which took as input the high-level information about the user who served as reward for the RL process. This could be, for example, user's impression about agent's W&C, or user's engagement. In this subsection we specify the RL methods which best fits our model and give the general formula that we implemented in our module. This formula could be adapted according to the different goals of the agent, as we did in the use case described in Chapter 8.

The general problem of RL can be abstracted by the framework of finite Markov decision processes (MDPs). MDPs are meant to be a straightforward framing of the problem of learning from interaction to achieve a goal. The framework depicted in Figure 7.4, can be adapted to our context:

- An *Agent*: it is the ECA that decides what behaviour to perform;

- An *Environment*: it is represented by information about user's reaction during the interaction with the agent;
- A *State* $S_t \in S$: it is a situation occurring at a time step t (e.g., user's impression), which is the basis on which the agent's choices are made;
- An *Action* $A_t \in A(s)$: it is the communicative intention of the agent. The choice of the action to perform is made by interacting with the user. The user responds to the action and presents new situations S_{t+1} to the agent;
- A *Reward* $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$: it is user's reaction to agent's behaviour. It is represented by a special numerical values that the agent seeks to maximize over time through its choice of actions.

The *agent*'s objective is to maximize the amount of *reward* it receives over time.

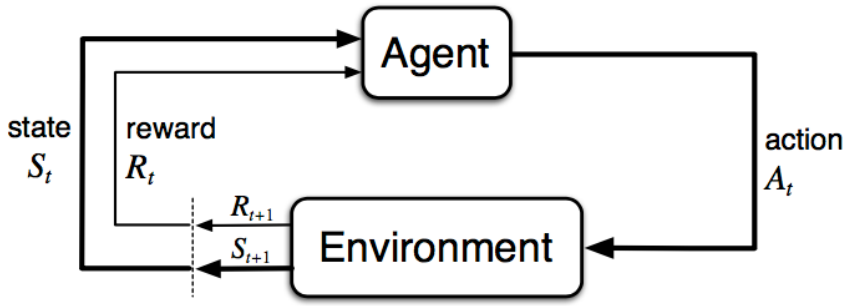


Figure 7.4 – The agent-environment interaction in a Markov decision problem. From Sutton and Barto (2018).

This framework may not be sufficient to represent all decision-learning problems usefully, but it has proved to be widely useful and applicable.

There exist three fundamental classes of methods for solving finite MDPs: dynamic programming, Monte Carlo methods, and temporal difference (TD) learning. In our context we searched for a model-free bootstrapping method. That is, a model that could learn directly from raw experience without requiring prior knowledge of the environment (model-free) and by updating the knowledge of the agent on every time-step (*action*) without waiting for a final outcome (bootstrapping). Monte Carlo methods are model-free but they update their knowledge at the final step. Dynamic programming methods update estimates online based on other learned estimates but they are not model-free. TD methods, instead, are model-free and bootstrapping, thus we selected this class of methods for our purposes.

In the general formula of TD, the value estimation V for a non-terminal state S at a time t is updated at each time step:

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)] \tag{7.1}$$

where $R_{t+1} + \gamma V(S_{t+1})$ is called TD target, that is, the estimated return of the value function.

In the equation two parameters that are important in RL are used:

- α: it is the *learning rate* and ranges in $[0, 1]$. It determines with which extension the new information acquired will overwrite the old information. A value near to 1 adjusts aggressively and would cause the agent to be interested only in recent information. A value near to 0 adjusts conservatively but would prevent the agent from learning.
- γ: it is the *discount rate* and ranges in $[0, 1]$. It determines the importance of future rewards. A value near to 0 will make the agent "opportunistic" by making sure that it only considers the current rewards. A value near to 1 will make the agent attentive even to the rewards he will receive in a long-term future.

One of the most used TD algorithms is Q-learning (Watkins, 1989) defined by the Bellman equation:

$$Q_{new}(s_t, a_t) = Q(s_t, a_t) + \alpha[R_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{7.2}$$

where:

- the value function Q does not concern only states but state-action combination;
- $\max_a Q(s_{t+1}, a)$ is the maximum expected future reward given the new state s_{t+1} and all possible actions a at that new state;
- $R_{t+1} + \gamma \max_a Q(s_{t+1}, a)$ is the learned value.

Q-learning is an off-policy method, that is, unlike on-policy methods like SARSA (Rummery and Niranjan, 1994), the action A_{t+1} is chosen in a greedy fashion without following a certain policy but by simply taking the max of Q over it.

As explained in Section 7.1, one of the challenges of all RL algorithms is the trade-off between exploration and exploitation, i.e., between getting new information and using current information. To face this dilemma, an ϵ-greedy policy is often applied, that consists in choosing the best action with $p = 1 - \epsilon$ and a random action with $p = \epsilon$, where ϵ can decrease over time to improve exploitation once getting enough information.

The output of our RL module was the specific verbal and non-verbal behaviour with which the agent would perform the dialog act selected by the *Dialog Planner*. The set of possible behaviours derived from our previous studies described in Chapter 5 and 6. The final communicative intention was coded into an FML file which was sent to the *Agent's animation module*.

7.5 Agent's Animation Module

This module took as input utterances enriched by nonverbal behaviour such as gestures, facial expressions, gaze, specified in FML-APML language (Mancini and Pelachaud, 2008).

The ECA's animation was realised by the Greta/VIB Platform (Pecune et al., 2014). It is a fully SAIBA compliant system (Vilhjálmsson et al., 2007) for the real-time generation and animation of ECA's verbal and nonverbal behaviours (see subsection 1.2.2). As described in the previous Section, we made use of the Dialog Manager Flipper to select the communicative intention of the agent. It represented the classic *Intent Planner* of SAIBA architecture.

The main components of the *Agent's Animation Module* were:

- *Behaviour Planner*, that transformed the communicative intents received in input into multi-modal signals;
- *Behaviour Realizer*, that produced the movements and rotations for the joints of the ECA;
- *MPEG4 Animatable*, that realized the face and body animation of the agent.

In our system, the FML file sent by the *Impressions Management Module*, which contained the communicative intention of the agent, was directly sent to the *Behaviour Planner* module. This module then transformed the communicative intention into synchronised multimodal behaviours, thanks to the information from the *Behaviour Lexicon*, containing the mappings between communicative intentions and multimodal behaviours. Finally, the *Behaviour Realizer* module instantiated the multimodal behaviours and the *MPEG4 Animatable* handled the synchronization with speech and generated the animations for the agent.

7.6 Conclusion

IN this Chapter we presented the general architecture for managing an ECA's impressions according to user's reactions. The system included 3 main modules. The first concerned the detection and interpretation of user's non-verbal behaviour (such as facial expressions, gaze, head and trunk rotation) and speech. The second module exploited a dialog manager and a reinforcement learning algorithm in order to select the communicative intention of the agent, composed by a dialog act performed by specific verbal and non-verbal behaviour, with the purpose of eliciting an impression of W&C. The last module concerned the realisation of this impression into an animation through the Greta/VIB Platform.

The key points of this Chapter:

Our computational model for an ECA impression management is composed by:

- the User's Analysis Module, which detects and interprets user's reactions through EyesWeb and machine learning tools;
- the Impressions Management Module, which exploits user's information to select the communicative intention of the agent, through a reinforcement learning algorithm and a dialog planner;
- the Agent's Animation Module, which allows for the generation of multimodal behaviours of the agent through the Greta/VIB Platform.

Chapter 8

User Study 1: Adapting agent's behaviour according to user's engagement

Contents

THIS Chapter presents our first use case where we applied the system architecture described in the previous Chapter. We were interested in investigating the link between user's engagement and user's impressions of agent's W&C. We adapted the modules of the architecture in order to compute user's engagement from the analysis of low-level signals, and to select the self-presentational intention realised by the agent in order to elicit different degrees of warmth and competence. An evaluation study is presented, where we applied our modified architecture in a real-time scenario where the agent played the role of a virtual guide of museum. In this study we compared the effects of an adapting agent and a non-adapting one on user's impressions and perception of the interaction.

8.1 Introduction

This Chapter presents our first application of the system architecture for ECA's impressions management in real-time described in the previous Chapter. The use case was an agent playing the role of a virtual guide of museum. Our goal was to manage W&C dimensions in order to obtain an engaging ECA, by following the idea that a more engaging agent is likely to form a positive impression and be accepted by the user, thus promoting further interactions (Bergmann et al., 2012; Cafaro et al., 2017). Other authors focused on different strategies to improve user's engagement, for example they focused on agent's backchannels, politeness strategies or verbal alignment (see Section 4.4). In this Thesis, since we focused on W&C impressions, we were interested on the role of these impressions on user's engagement, in particular whether adapting the agent's W&C impressions could affect user's engagement.

According to this reasoning, we focused on two main research questions:

- (Q1) *Is there a relationship between agent's impressions of W&C and user's engagement during the interaction with an ECA?*
- (Q2) *Is it possible to improve user's engagement by managing agent's degree of W&C?*

In order to answer these questions, we focused on the effects of self-presentational strategies that could be performed by the agent in order to manage its impressions of W&C. For example, the agent could decide to present itself as a warm guide, or to highlight its level of competence by decreasing its warmth. At the beginning of the interaction the agent had no information about the effects of its strategy on the user's perception. It could use user's engagement level as a measure to assess which strategy worked better, that is, which strategy increased user's engagement.

We customised the module *User's Analysis* (see Section 7.3) in order to compute user's engagement from low-level signals and use it as reward for the RL algorithm. The module *Impressions Management* (see Section 7.4) was also adapted by including a self-presentational intention planner.

This Chapter is organised as follows: the next Section contains a brief overview of engagement in human-agent interaction; in Section 8.3 the 4 self-presentational strategies used by the agent in this use case are presented; Section 8.4 describes the modules of the architecture that have been modified to adapt the model to our goals; in Section 8.5 we present an evaluation study where we investigated the effects on user-agent interaction of an adapting agent compared to a non-adapting one.

8.2 Engagement in Human-Agent Interaction

Despite of being a major theme of research and a universal goal in Human-Computer Interaction (HCI), engagement is a difficult concept to define (102 different definitions of engagement exist according to Doherty & Doherty review (Doherty and Doherty, 2018)), due to its multidimensional nature and the difficulty to measure it.

A detailed summary of engagement definitions in human-agent interaction is provided in Glas and Pelachaud (2015a). Among others, it can be defined as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction" (Poggi, 2007), and as "the process by which participants involved in an interaction start, maintain and terminate an interaction" (Corrigan et al., 2016; Sidner and Dzikovska, 2005).

Engagement is not measured from single cues, but rather from several cues that arise over a certain time window (Peters et al., 2005b). Engagement can be defined by high-level behaviour like: synchrony – which is the temporal coordination during social interactions; mimicry – which is the automatic tendency to imitate others; feedback – which can indicate whether the communication is successful or not. Similarly, engagement can also be defined by low-level behaviour like: eye gaze - providing feedback and showing interest; head movements - nods (in agreement, disagreement, in between); gestures - to greet, to take turns; postures - body orientation, lean; facial expressions.

Several authors attempted to design engaging virtual agents, by focusing on the use of feedback and backchannels (Truong et al., 2010), by adopting politeness strategies (Glas and Pelachaud, 2015b), or by investigating the role of verbal alignment for improving user's engagement (Campano et al., 2015b) (see Section 4.4). Clavel et al. (2016) provided a review on methodologies for assessing user's engagement in human-agent interaction. Other studies focused on how to improve user's engagement by adapting social agents (mainly robots) behaviours, using RL methods. These works incorporated user's social signals to measure user's engagement and exploited it as the reward of the RL algorithm. For example, Ritschel et al. (2017) computed user's engagement as a reward, with

the goal to adapt robot's personality expressed by linguistic style. Gordon et al. (2016) exploited facial expressions to measure child's engagement in order to adapt a robot's behaviours, while Liu et al. (2008) exploited user's physiological signals.

In the work presented in this Chapter we used low-level signals, such as facial Action Units activation, trunk and head rotation, to measure engagement. The engagement detection model is described in subsection 8.4.1.

8.3 Self-presentational strategies

Jones and Pittman (1982) argued that people can use different verbal and non-verbal behavioural techniques to create the impressions they desired in their interlocutor. The authors proposed a taxonomy of these techniques, that they called self-presentational strategies. We illustrate here 4 of their strategies that can be associated to different levels of W&C. We did not consider the 5th strategy of the taxonomy, called *Exemplification*. This strategy is used when people want to be perceived as self-sacrificing and to gain the attribution of dedication from others, thus it is not related neither to warmth nor to competence. Concerning the other 4 strategies, two of them focus on one dimension at a time, the other two focus on both dimensions by giving them opposite values:

- *Ingratiation*: its goal is to get the other person to like you and attribute positive interpersonal qualities (e.g., warmth and kindness). In our case, the agent selecting this strategy had the goal to elicit impressions of high warmth towards the user, without considering its level of competence.

- *Supplication*: it occurs when individuals present their weaknesses or deficiencies to receive compassion and assistance from others. In our case, the agent selecting this strategy had the goal to elicit impressions of high warmth and low competence.

- *Self-promotion*: it occurs when individuals call attention to their accomplishments to be perceived as capable by observers. In our case, the agent selecting this strategy had the goal to elicit impressions of high competence, without considering its level of warmth.

- *Intimidation*: it is defined as the attempt to project one's own power or ability to punish to be viewed as dangerous and powerful. In the context of our research, we interpreted this strategy in a smoother way, as the goal to elicit impressions of low warmth and high competence.

In our use case, for each speaking turn, the agent played one out of these 4 self-presentational techniques.

These techniques were realised by the ECA through its verbal and non-verbal behaviour. The verbal behaviour characterizing the different strategies was inspired to the

works of Pennebaker (2011) and Callejas et al. (2014). According to their findings, we manipulated the use of *you* and *we* pronouns, the level of formality of the language, the length of the sentences. For example, sentences aiming at eliciting high warmth contained more pronouns, less synonyms, more informal language, so that the phrases would be more casual and would give the impression to be less meditated; more verbs rather than nouns, and positive contents were predominant. Sentences aiming at eliciting low warmth contained more negations, longer phrases, formal language, and did not refer to the speaker. Sentences aiming at eliciting high competence contained high rates of we- and you-words, and I-words at low rates. Figure 8.1 shows the use of verbal behaviour according to the different levels of W&C on their two axes, while Table 8.1 shows an example of an agent's utterance for each of the 4 self-presentational techniques.

Strategy	Translated sentence	Original sentence
Ingratiation	"You can test some games, if you wanna."	*Tu vas pouvoir tester des jeux si tu veux.*
Supplication	"I dunno about the other exhibits of the museum, but here you can test some games, it's cool!"	*J'connais pas les autres expo du musée, mais ici on peut tester des jeux, c'est trop bien !*
Self-promotion	"In this exhibit, you can test some videogames."	*Dans cette expo tu vas pouvoir tester des jeux-vidéos.*
Intimidation	"In this exhibit, you can try out some games on different platforms."	*Dans cette exposition tu peux essayer des jeux sur différents supports.*

Table 8.1 – An example of 4 different sentences for the same dialog act (the agent introduces the video games exhibit), according to the 4 different self-presentational techniques. The original sentences in French are provided.

The choice of agent's non-verbal behaviour was based on our previous studies described in Chapter 5 and 6. In particular, we manipulated the type of gestures, the type of arms rest poses and smiling behaviour.

So, for example, if the current agent's self-presentational strategy was *Supplication* and the next dialog act to be spoken was introducing a topic, then the agent would say "I think that while you play there are captors that measure tons of stuffs!" accompanied by smiling and beat gesture. Conversely, if the current agent's self-presentational strategy was *Intimidation* and the next dialog act to be spoken was the same, then the agent would say "While you play at videogames, several captors measure your physiological signals." accompanied by no smiling and ideational gesture.

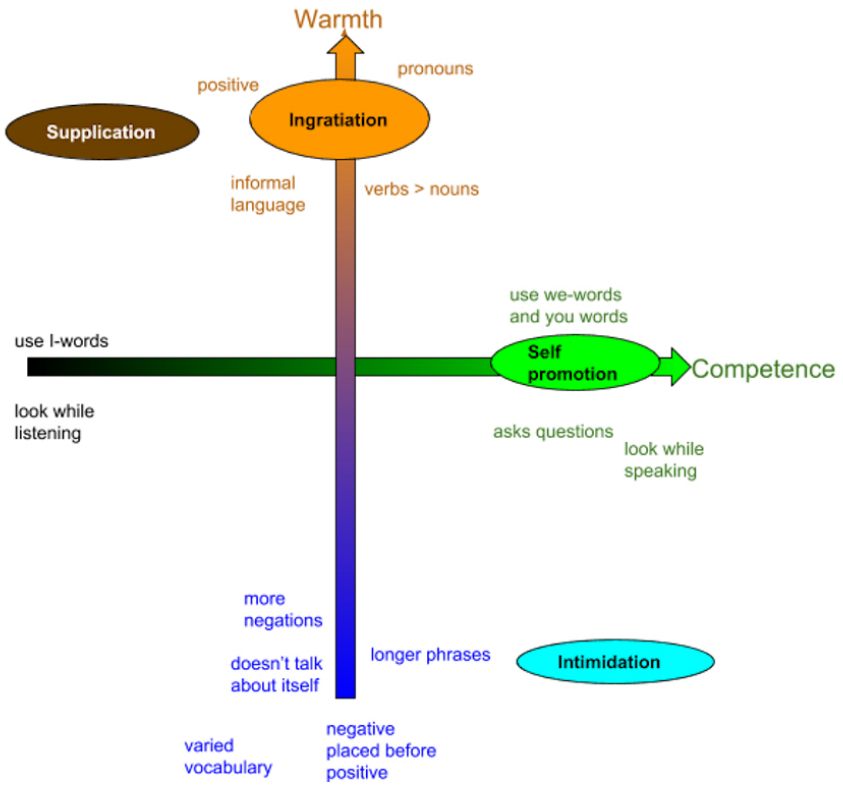


Figure 8.1 – Use of pronouns, verbs, type of language, and other verbal behaviours associated to each self-presentational technique, inspired from (Pennebaker, 2011) and (Callejas et al., 2014) works.

8.4 System Architecture

As mentioned at the beginning of this Chapter, we customised the general architecture described in Chapter 7 in order to better fit our purpose, that is to select agent's self-presentational techniques to maximise user's engagement. The modified architecture of the system is depicted in Figure 8.2. In the following Section we give more details about the modified modules.

8.4.1 Engagement Fusion module

Overall user's engagement was computed continuously at the end of every speaking turn, defined as the time window from the moment when the agent started to pronounce its question for the user to the moment when the user stopped replying to the agent (or, if the user did not respond, until a threshold of continuous silence was reached). After the end of the speaking turn, the overall mean engagement was sent from EyesWeb to the RL module, described in the following section, that selected the self-presentational technique the ECA would use in the next speaking turn.

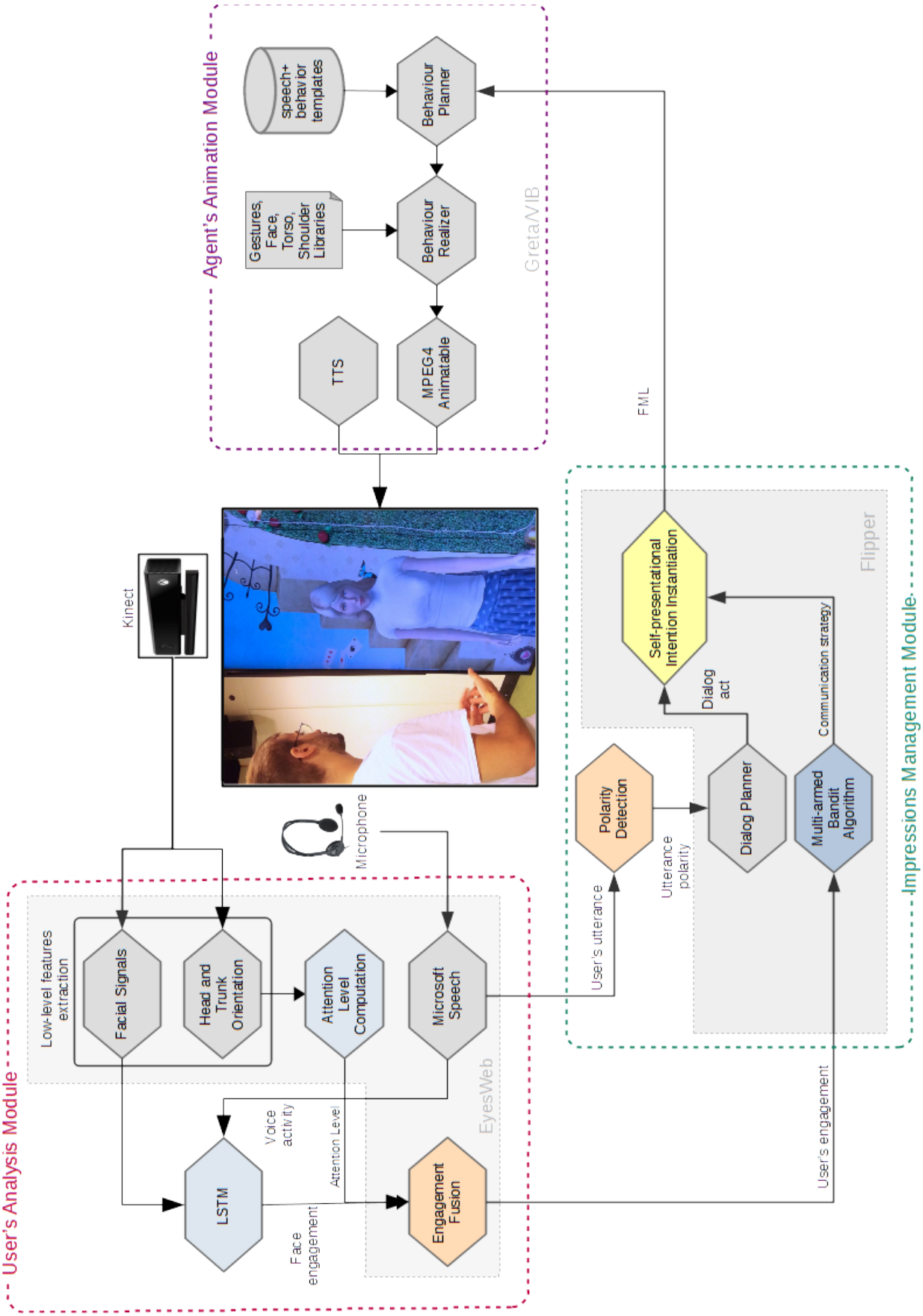


Figure 8.2 – The modified version of the general architecture described in Chapter 7. The modules that have been modified are coloured, while the modules that did not change are grey. In particular, the *User's Analysis Module* was customised in order to compute users' engagement. The *Impressions Management Module* was also modified by implemented a *Self-presentational Intention Instantiation* module and by adapting the reinforcement learning algorithm to our use case.

The computational model of user's engagement was based on the detection of multimodal signals that, according to the literature described in subsection 8.4.1, can be indicators of engagement. In particular, we addressed two main classes of non-verbal signals:

- *Facial signals* - Similar to other works like Castellano et al. (2009), Corrigan et al. (2016), we aimed to detect facial signals to quantify user's engagement. Smiling is usually considered an indicator of engagement, as it may show that the user is enjoying the interaction (Castellano et al., 2009). Eyebrows are equally important: for example, Corrigan et al. (2016) claimed that "frowning may indicate effortful processing suggesting high levels of cognitive engagement". Subsection 8.4.1.1 presents our model of engagement detection from facial signals.

- *Head/trunk signals* - According to Corrigan et al. (2016), attention is a key aspect of engagement: an engaged user continuously gazes at relevant objects/persons during the interaction: the longer her attention is focused, the more engaged she is. Conversely, according to Sidner and Dzikovska (2005) and Peters et al. (2008), "turning one's body away from the other participant" or "looking away" can indicate disengagement, and "engagement may be diminished due to not engaging in shared attention behaviour". We approximated user's gaze with the user's head and trunk orientation, as reliable eye tracking would require invasive hardware or introduce too many constraints on user's movement (e.g., the user had to wear a glass-mounted eye tracker or sit very close to a sensor). Subsection 8.4.1.2 presents our model of attention computation from head/trunk signals.

In the following subsections we illustrate how we implemented our model by describing how we extracted face, head and trunk signals and how we computed the user's engagement. We exploited machine learning techniques to extract user' engagement from face Action Units: we trained a Long Short-Term Memory model (subsection 8.4.1.1) with annotated human one-to-one interaction data. Then, we computed user's attention by measuring how long the user's head and trunk were facing the agent during the speaking turn (subsection 8.4.1.2). The attention level was finally added or subtracted to the engagement previously detected on the face to compute the final level of user's engagement (subsection 8.4.1.3).

8.4.1.1 Engagement Detection from Facial Signals[1]

In order to detect user's engagement from facial signals, we applied the model developed by Dermouche and Pelachaud (2018). This model allowed measuring the evolution of engagement over time and was found to perform better than other models, as shown in Table 8.2.

[1]This work has been realised by Soumia Dermouche, CNRS-ISIR, Sorbonne University, Paris.

Model	Parameters	Recall	Precision	F-measure
Random	uniform probability over classes	19%	18%	19%
Naive Bayes	-	39%	39%	37%
Random Forest	10 trees, max. tree depth=5	49%	50%	48%
AdaBoost	-	50%	51%	50%
Decision Tree	max. tree depth=5	50%	48%	47%
Neural Net	multilayer perceptron, alpha=1	68%	68%	68%
Our model (LSTM)	-	77%	76%	76%

Table 8.2 – Prediction of expert engagement based on different models. The LSTM model of Dermouche and Pelachaud (2018) (last row) performs better than the other methods.

The engagement detection model from facial signals was based on a Long Short-Term Memory (LSTM) prediction model using Recurrent Neural Networks implemented with the Keras toolkit and TensorFlow. The LSTM model was trained on the French part of NoXi dataset (the same described in Chapter 5). More details about the model are available in Dermouche and Pelachaud (2018).

The module took as input user's AUs and the conversational state of the interaction during the last 30 frames. In particular, the intensity (from 0 to 5) of 17 AUs (*AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU45*), the presence or activation (0 absent, 1 present) of 18 AUs (*AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28, AU45*) and the conversational state of the interaction (none, both, user or agent is speaking). These inputs were detected in real time using EyesWeb.

The output of the model, at the end of each speaking turn, was the user's engagement level E, a number in $[1, 5]$, where 1 meant that the user was strongly disengaged, and 5 meant that the user was strongly engaged.

8.4.1.2 Attention Detection from Head/Trunk Signals[2]

To obtain a more accurate user's level of engagement, we detected user's attention level Peters et al. (2005a) and we fused this level with the engagement level described just above (subsection 8.4.1.1).

Attention level computation was implemented by Professor Maurizio Mancini in EyesWeb as a set of rules. It took as input user's head and trunk orientation and computed user's attention level. For example, if the user was looking at the agent (with both her face and her trunk orientations) then the attention level increased and we applied a bonus to the final score of engagement.

[2]This work was realised in collaboration with Prof. Maurizio Mancini

While we could easily access user's head orientation extracted by the Kinect, we needed some processing to compute the trunk orientation. As illustrated in Figure 8.3, the agent was displayed on a large screen on top of which the Kinect camera was attached. We extracted the user's trunk 3D rotation angle by computing the following straight lines:

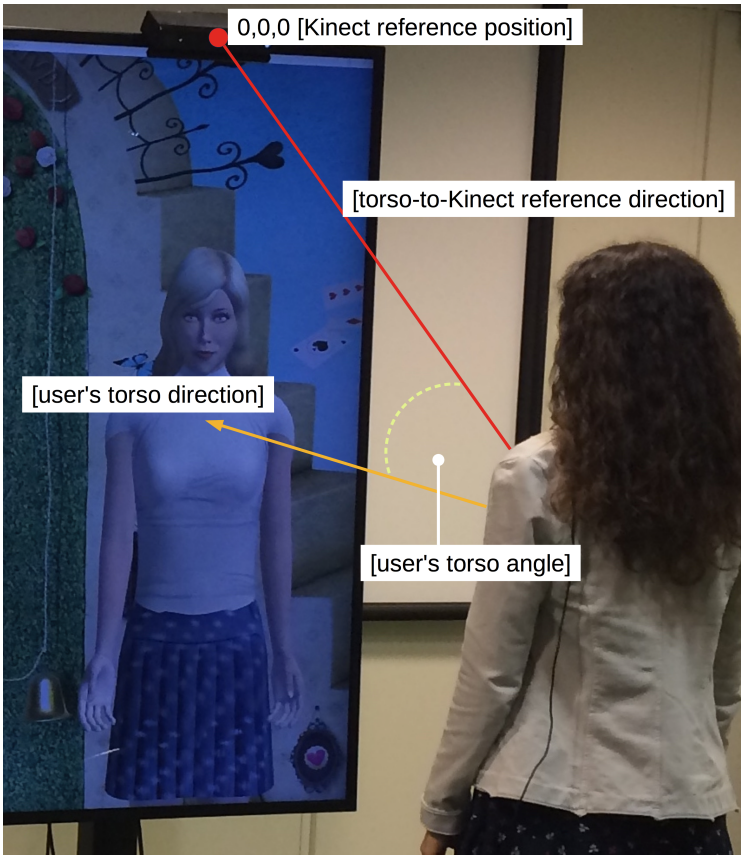


Figure 8.3 – User's trunk orientation computation. We extracted the angle between the 3D orientation of user's trunk (yellow line) and a reference direction (the red line between the user's trunk and the Kinect sensor).

- $L(K, U_T)$ - the line between the Kinect K (that was always located at the 3D space origin) and the user's trunk U_T. It is the red line in Figure 8.3. We took the geometric center of user's shoulders as reference for the user's trunk position;
- $L(U_{LS}, U_{RS})$ - the line between the user's left U_{LS} and right U_{RS} shoulders;
- L_{U_T} - the line that was orthogonal to $L(U_{LS}, U_{RS})$ (the yellow line in Figure 8.3).

The user's trunk orientation was equal to the angle (in the 3D space) between $L(K, U_T)$ and L_{U_T}. In this way, the trunk orientation was relative to the Kinect position (the same happened for the head rotation extracted by the Kinect) and did not depend on the user position. That is, even if the user was not standing exactly in front of the screen/camera

but she was on the side of the screen, the rotations would be relative to the Kinect and, consequently, to the agent position.

To compute the final values of head and trunk orientations we performed a calibration phase, in which we recorded the rotation values of a person whose head and trunk were oriented toward and away (i.e., turning far on the left and on the right) from the agent. For the evaluation study presented in Section 8.5, we calibrated the system only one time, after installing it in the experiment room, since the processing we performed was not user-dependent but it was influenced only by the physical location of the screen and Kinect.

The final rotation values were normalized depending on the calibration values, so that a head/trunk orientation value of 1 meant "toward the agent" while 0 meant "away from the agent".

At the end of each speaking turn, to compute user's attention level A, we took into account her head/trunk orientation over time for the duration of the speaking turn. If A was low (i.e., toward 0) it meant that the user did not look at the agent most of the time the agent spoke; if it was high (i.e., toward 1) it meant that the user looked at the agent most of the speaking turn time.

8.4.1.3 Engagement Fusion

Once engagement E had been detected from user's facial signals and attention level A had been computed from user's head/trunk orientation, we fused them to obtain the final value of user's engagement E_f. To do that, based on studies showing that head/trunk orientation contribute to indicate user's engagement or disengagement (Sidner and Dzikovska, 2005), the value of A was used as a bonus/malus to modulate the value of E.

8.4.2 Self-presentational Intention Instantiation

During its interaction with the user, the ECA had the goal of selecting its *self-presentational intention* (e.g., to communicate verbally and non-verbally a given dialog act with high warmth and low competence). The ECA could choose its intention among a given set of possible utterances depending on the user's overall engagement value. For example, if the last self-presentational intention had the effect of decreasing the detected user's engagement, then the ECA would select a different intention for the next speaking turn, that is, it would select a self-presentational technique associated with a different value of W&C; conversely, if the last intention increased user's engagement, that intention would be maintained.

This problem can be seen as a *multi-armed bandit problem* (Katehakis and Veinott Jr, 1987), a simplified setting of RL which models agents evolving in an environment where they can perform several actions, each action being more or less rewarding for them.

In our case, the *actions* that the ECA could perform were the verbal and non-verbal behaviours corresponding to the self-presentational intention the ECA aimed to commu-

nicate, and they were selected by the equation 8.1. The *environment* was the interaction with the user, while the *state* space was the set of dialog acts used at each speaking turn, and it was defined by the Dialog Planner. The choice of the *action* did not change the *state* (i.e., the dialog act used during the actual speaking turn), but rather it acted on how this dialog act was realized by verbal and non-verbal behaviour.

In order to maximize user's engagement during the interaction, the ECA, at the beginning, explored the environment (i.e., by randomly choosing an initial self-presentational intention) and then exploited its knowledge (i.e., user's engagement) to find the most rewarding self-presentational intention.

To do that, we chose to exploit the ϵ-decreasing learning approach (see subsection 7.4.2): the exploration rate ϵ continuously decreased in time. In this way, the ECA started the interaction with the user by exploring the environment without taking into account knowledge (i.e., user's engagement) and finished it by exploiting its knowledge only (i.e., without performing any further environment exploration). That is, the ECA explored with probability ϵ, and exploited knowledge with probability $1 - \epsilon$.

The ECA updated its knowledge through a table where it iteratively approximated the expected reward $Q(int)$ of a self-presentation intention int. This was done using the formula:

$$Q(int)_{t+1} \leftarrow (1 - \alpha) \times Q(int)_t + \alpha \times e_t \tag{8.1}$$

where:

- $Q(int)$ was the expected value of the self-presentational intention, $int \in$ [*ingratiation, supplication, self-promotion, intimidation*];
- α was the learning rate, set at 0.5, a very high number compared to other works (e.g., Burda et al. (2018) set it to 0.0001). This was because the ECA needed to learn quickly (i.e., in few dialogue steps) the self-presentational intention to use;
- e was the overall engagement score, that is the reward for the ECA.

8.5 User Study

We now present the experimental study we conceived to investigate whether or not an ECA endowed with the architecture described in the previous section, that is, able to manage its impressions of W&C according to user's engagement, could affect user-agent interaction. In the study, we compared different conditions where the ECA could interact with the user by adapting or not its behaviours.

We created a scenario where the virtual agent, called Alice, played the role of a virtual guide of a museum. The experiment took place at the Carrefour Numérique, an area of the Cité des sciences et de l'industrie in Paris, one of the largest sciences museums in Europe.

8.5.1 Independent Variables

The independent variable manipulated in this study concerned agent's *Strategy*, that is, how the agent managed its behaviours to influence user's perception of its W&C. According to the different *Strategy* conditions, the agent could select one of the 4 self-presentational techniques at the start of the interaction and display it during the whole interaction, or select one of the 4 at each speaking turn, either randomly or by using our self-presentational intention model based on user's overall engagement detection.

In total, *Strategy* had 6 levels:

- **INGR**: when the agent selected the Ingratiation self-presentational technique from the beginning to the end of the interaction, without considering user's reactions;
- **SUPP**: when the agent selected the Supplication self-presentational technique from the beginning to the end of the interaction, without considering user's reactions;
- **SELF**: when the agent selected the Self-promotion self-presentational technique from the beginning to the end of the interaction, without considering user's reactions;
- **INTIM**: when the agent selected the Intimidation self-presentational technique from the beginning to the end of the interaction, without considering user's reactions;
- **RAND**: it consisted in selecting one of the 4 self-presentational techniques, randomly, at each speaking turn, without considering user's reactions;
- **IMPR**: it consisted in selecting one of the 4 self-presentational techniques, at each speaking turn, by using our self-presentational intention model based on user's overall engagement detection (see subsection 8.4.1).

According to *Strategy* level, the multi-armed bandit algorithm was applied (or not) to update the action (i.e., the following self-presentational intention) of the agent. The choice of the condition was made by modifying initial settings of the *Impressions Management Module*.

8.5.2 NARS

Before the interaction, we collected information about users' attitudes and prejudices towards virtual characters. We used a slightly adapted version of the Negative Attitudes towards Robots Scale (*Nars*) from Nomura et al. (2006). This questionnaire measures people's a-priori negative attitudes toward situations and interactions with robots, toward the social influence of robots, and toward emotions in interaction with robots. We selected the most relevant items according to our context and adapted the questions by referring to virtual characters instead of robots. Participants gave their ratings on a 5-points Likert

scale, from 1 = "I completely disagree" to 5 = "I completely agree". The items of the questionnaires (translated in English) are available in Table 8.3. In our analyses we checked if *Nars* scores had an impact on the dependent variables (described in the next paragraph).

Items
1. I would feel uneasy if virtual characters had emotions.
2. I would feel relaxed talking with virtual characters.
3. I feel comforted being with virtual characters that have emotions.
4. The word "virtual character" means nothing to me.
5. I would hate the idea that virtual characters were making judgements about things.
6. I would feel very nervous just standing in front of a virtual character.
7. I would feel paranoid talking with a virtual character.
8. I am concerned that virtual characters would be a bad influence on children.

Table 8.3 – Items of the *Nars* questionnaire, adapted from Nomura et al. (2006).

8.5.3 Dependent Variables

The dependent variables were measured during and after the interaction with the ECA. During the interaction, if the participant agreed in the consent form, we recorded the users speech audio, in order to measure users' cues of engagement from their verbal behaviour. After the interaction with the ECA, participants were asked to fill in some questionnaires where they were asked to rate the agent's W&C and their overall satisfaction of the interaction.

8.5.3.1 Verbal cues of engagement

For people who agreed with audio recording of the experiment, we collected quantitative information about their answers, in particular:

- The polarity of the answer to Topic1_question (see subsection 8.5.5);
- The polarity of the answer to Topic2_question (see subsection 8.5.5);
- The number of any verbal feedback produced by the user during a speaking turn.

8.5.3.2 Self-report assessment

After the interaction, the participants filled in a final questionnaire, divided in several parts. In particular we measured:

- User's perception of agent's warmth (*w*) and competence (*c*): we presented a list of adjectives referring to W&C and asked participants to indicate their agreement on a 5-points Likert scale (1 = "I completely disagree", 5 = "I completely agree") about how precisely each adjective described the character. The items were taken from Aragonés et al. (2015) scale, and were: *kind, pleasant, friendly, warm* for warmth, and *competent, effective, skilled, intelligent* for competence.

- User's perception of the interaction (*perception*): the second part of the questionnaire concerned a list of items adapted from those already used by Bickmore et al. (2011). They are shown in Table 8.4.

Measure	**Question**
satisfaction	*I am satisfied with my interaction with Alice.*
continue	*I would like to talk with Alice again.*
like	*I liked Alice.*
learnfrom	*I have learned something from Alice.*
expo	*Alice made me want to visit the exposition (if you haven't yet)*
rship	*I would describe Alice as a complete stranger vs a close friend.*
likeperson	*I would describe Alice just as a computer vs like a person.*

Table 8.4 – Items of the questionnaire about user's perception of the interaction, adapted from Bickmore et al. (2011). Alice was the name of the virtual character.

8.5.4 Hypotheses

With this study we firstly aimed to investigate if the ECA's 4 self-presentational techniques during all the interaction were correctly perceived by users, for example, if users rated the agent in **INGR** condition as warm, and the agent in **INTIM** as cold and competent.

In line with this goal, we hypothesised that:

- **H1ingr**: The agent in **INGR** condition would be perceived as *warm* by users;

- **H1supp**: The agent in **SUPP** condition would be perceived as *warm* and *not competent* by users;

- **H1self**: The agent in **SELF** condition would be perceived as *competent* by users;

- **H1intim**: The agent in **INTIM** condition would be perceived as *competent* and *not warm* by users.

Then, our main hypothesis was that the use of the our adapting model based on user's overall engagement detection (i.e., when the virtual character adapted its behaviours) positively affected user's perception of the interaction. Thus we hypothesised that:

- **H2a**: The scores of *perception* items would be higher in **IMPR** condition compared to all the other strategies;
- **H2b**: The agent in **IMPR** condition would influence how it was perceived in terms of W&C.

8.5.5 Protocol

The experiment took place in a room of the Carrefour Numérique. As shown in Figure 8.4, the room was divided into three areas:

- the questionnaires place, including a desk with a laptop and a chair;
- the interaction place, with a big screen displaying the virtual character, a Kinect 2 on the top of the screen and a black tent in front of the screen;
- the control station, separated from the rest of the room by 2 screens. This place included a desk with the computer running the system architecture.

The experiment was completed in three phases:

1. Before the interaction begun, the participant sat at the questionnaires place, read and signed the consent form, and filled in a first questionnaire (see subsection 8.5.2), then moved to the interaction place, where the experimenter gave the last instructions [5 min];
2. During the interaction phase, the participant stayed right in front of the screen, between it and the black tent. He/she wore a headset and was free to interact with the virtual character as he/she wanted. During this phase, the experimenter stayed in the control place, behind the screens [3 min];
3. After the interaction, the participant came back to the questionnaires place and filled in the last questionnaires (see Section 8.5.3.2). After that, the experimenter proceeded with the debriefing.

The interaction with the ECA lasted about 3 minutes. It included 25 to 36 steps, according to user's answers. A step included one dialog act played by the agent and user's answer. If user did not reply in a certain interval of time, the agent continued with the following step. After each speaking turn, user's engagement was computed through our fusion engagement module (see subsection 8.4.1).

The dialogue was divided into 4 main parts that were always played by the agent, no matter what answers the users gave:

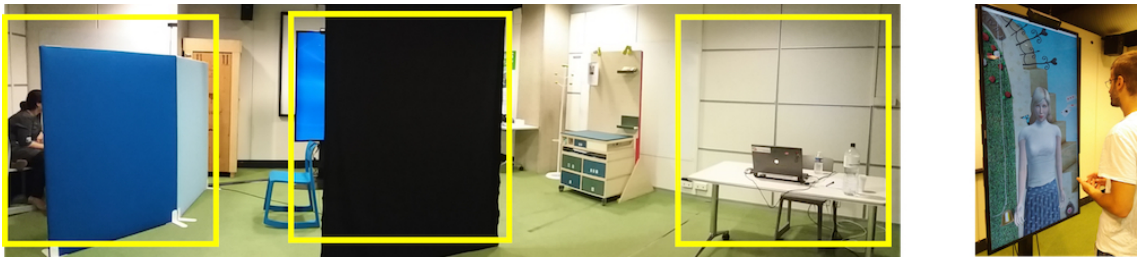


Figure 8.4 – The experiment room and an example of an interaction. In the yellow squares, on the left, the control place, in the middle the interaction place, and on the right the questionnaires space.

1. *Start interaction* (8 steps);
2. *Topic 1* (3 steps);
3. *Topic 2* (4 steps);
4. *End of the interaction* (4 steps).

At the end of parts *Start interaction*, *Topic 1* and *Topic 2*, Alice asked a question to the user. The Videogames_question asked if the user liked videogames. The Topic1_question asked if the user wanted to continue to discuss about Topic 1. The Topic2_question asked if the user wanted to discuss about Topic 2. After Topic1_question and Topic2_question, if the user gave a positive answer, the agent continued to talk about the same topic (6 steps for Topic 1, 5 steps for Topic 2), otherwise it skipped to the next part. The dialogue flowchart is shown in Figure 8.5.

8.5.6 Analysis and Results

We analysed data from 75 participants, of which 30 females and 2 preferring not to specify their gender. The majority of the participants were in the 18-25 or 36-45 age range, 5 of them were not native French speakers (but spoke and understood French), and 72% of them had at least a Bachelor. Participants were almost equally distributed across the levels of the independent variable *Strategy* (12.5±1 participants per each strategy).

Before conducting our analyses, we computed Cronbach's alphas for *w* and *c* items respectively and explored the distribution of data. Good reliability for *w* and *c* items was found ($\alpha = 0.9$ and $\alpha = 0.8$ respectively). We then grouped the 4 adjectives measuring *w* and the 4 measuring *c* and used the mean of these grouped items for each participant for our analyses. We ran ANOVAs on these data since their distribution satisfied assumptions for this test.

Nars scores got an acceptable score of reliability ($\alpha = 0.66$), we therefore computed the means of these items in order to obtain one overall mean for each participant. We then divided participants into 2 groups, "high" and "low", according to whether they obtained

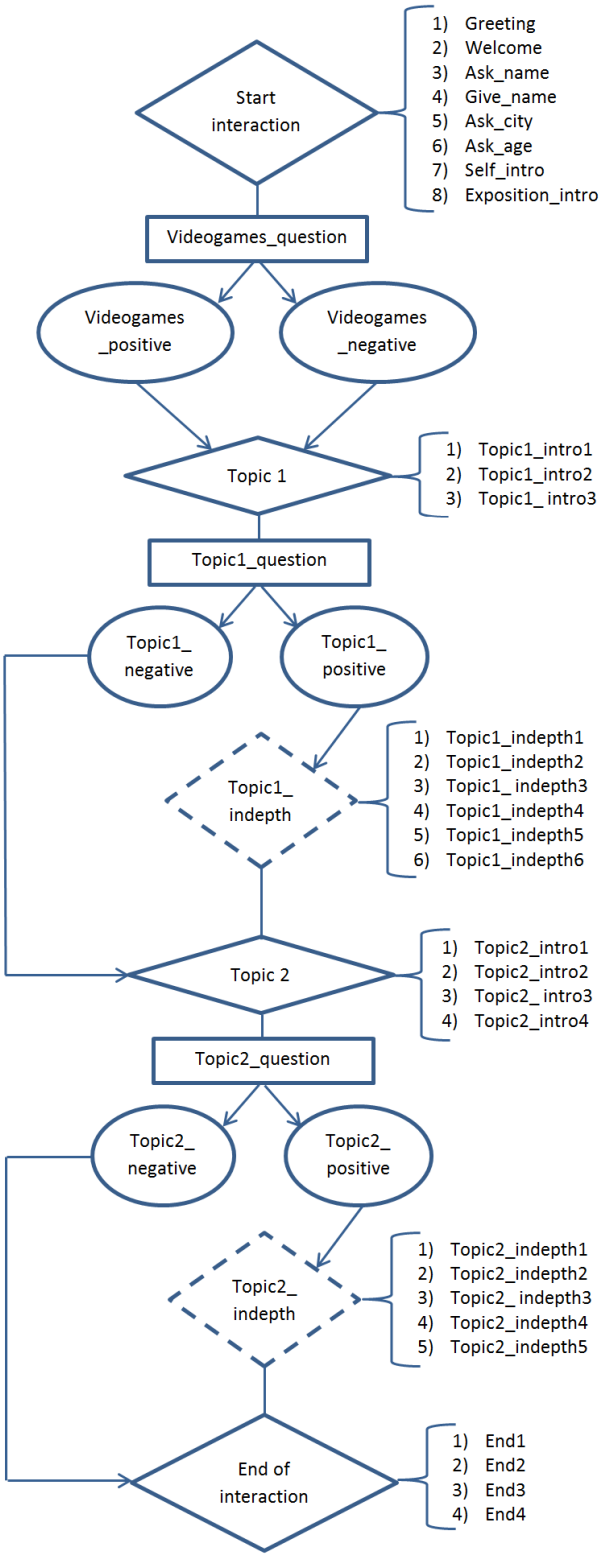


Figure 8.5 – The dialogue flowchart. The diamond shapes represent the main parts that always occurred during the dialogue, the rectangles represent questions, the rounds represent agent's reply to user's answer and the dotted shapes the optional parts. Where not specified, each shape represents one step of the dialogue.

Condition	Warmth mean $\pm$ SD
INGR	3.77 ± 0.57
SUPP	3.54 ± 0.999
SELF	3.81 ± 0.70
INTIM	2.63 ± 0.93
RAND	3.71 ± 0.80
IMPR	3.89 ± 0.38

Table 8.5 – Mean and standard deviation of warmth scores for each level of *Strategy*.

a score higher than the overall mean or not, respectively. Participants were almost equally distributed into the two groups (39 in the "high" group, 36 in the "low" group, almost equally distributed across the other variables, too).

8.5.6.1 Warmth

A 4-way between-subjects ANOVA, including age, sex, *Nars* scores and *Strategy* as factors, was first run in order to check for any effect of these variables. No effect of age and sex was found, so we then conducted a 4x2 between-subjects ANOVA with *Strategy* and *Nars* as factors. The analysis revealed a main effect of *Strategy* ($F(5, 62) = 4.75$, $p = 0.000974$, $\eta^2 = 0.26$) and *Nars* ($F(1, 62) = 5.74$, $p = 0.02$, $\eta^2 = 0.06$). Post-hoc test specified that *w* ratings were higher from participants in the "high" *Nars* group ($M = 3.74$, $SD = 0.77$) than from those in the "low" *Nars* group ($M = 3.33$, $SD = 0.92$).

Table 8.5 shows mean and SD of *w* scores for each level of Condition. Multiple comparisons t-test using Holm's correction showed that the *w* mean for **INTIM** was significantly lower than each of all the others (see Table 8.5). As consequence, the others conditions were rated as warmer than **INTIM**. **H1ingr**, **H1supp** were thus validated, and **H1intim** and **H2b** were validated for the warmth component.

8.5.6.2 Competence

A 4-way between-subjects ANOVA, including age, sex, *Nars* scores and *Strategy* as factors, was first run in order to check for any effect of these variables. No effects for any factor were found. When looking at the means of *c* for each condition (see Table 8.6), **SUPP** was the one with lower score, even if its difference with the other scores did not reach statistically significance (all p-values > 0.1). **H1supp** and **H1intim** (for the competence component) were not validated.

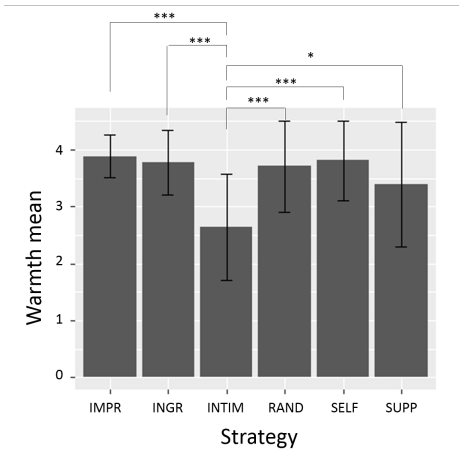


Figure 8.6 – Mean and SD values of warmth ratings for each level of *Strategy*. **INTIM** scores were significantly lower than each of any other condition. Significance levels: $*$: $p < 0.05, ** : p < 0.01, *** : p < 0.005$.

Condition	**Competence mean** $\pm$ SD
INGR	3.6 ± 0.62
SUPP	2.98 ± 0.77
SELF	3.75 ± 0.63
INTIM	3.65 ± 0.79
RAND	3.5 ± 0.70
IMPR	3.43 ± 0.76

Table 8.6 – Mean and standard deviation of competence scores for each level of *Strategy*. No significant differences among the conditions were found.

8.5.6.3 User's perception of the interaction

We analysed each item of *perception* separately, by applying non-parametric tests since data were not normally distributed.

Concerning *satisfaction* scores, a Kruskal-Wallis rank test showed a statistically significant difference according to *Strategy* ($H(5) = 11.99$, $p = 0.03$). In particular, Dunn's test for multiple comparisons found that **INGR** scores were significantly higher than **SUPP** ($z = 2.88$, p-adj $= 0.03$) and **INTIM** ($z = 2.56$, p-adj$= 0.04$). No differences were found between **IMPR** scores and the other conditions. In addition, a statistically significant difference between scores was found according to *Nars* scores ($U = 910.5$, $p = 0.02$): participants who got high scores in the *Nars* questionnaire were more satisfied by the interaction ($M = 3.62$, $SD = 0.94$) than people who got low scores in the *Nars* questionnaire ($M = 3.00$, $SD = 1.07$). Another interesting result concerned the effect of age on

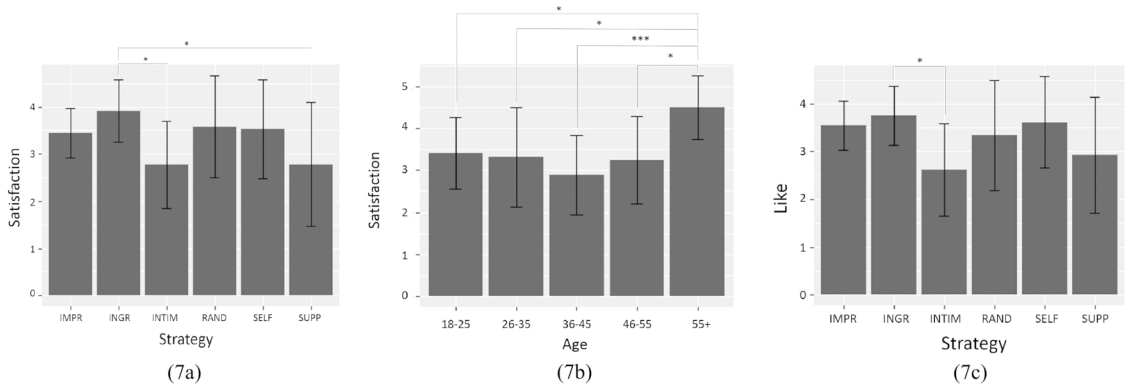


Figure 8.7 – Mean values with SD for the different items of *perception* where an effect of *Strategy* and age was found. Significant results of Dunn's test for multiple comparisons are reported, with the following significance levels: $* : p < 0.05, ** : p < 0.01, *** : p < 0.001$. (7a) mean values of *satisfaction* for each level of *Strategy*; (7b) mean values of *satisfaction* for each age range; (7c) mean values of *like* for each level of *Strategy*.

satisfaction ($H(4) = 15.05$, $p = 0.005$): people in the age range 55+ were more satisfied than people of each of any other age range (all p-adj ≤ 0.03).

Concerning *continue* scores, no effect of *Strategy* was found. In general, mean scores were not very high, with only scores in **INGR** and **SELF** conditions being higher than 3. A Mann-Whitney U Test showed a statistically significant difference according to *Nars* scores ($U = 998$, $p = 0.001$): participants who got high scores in the *Nars* questionnaire were more motivated to continue the interaction ($M = 3.28$, $SD = 1.12$) than people who got low scores in the *Nars* questionnaire ($M = 2.36$, $SD = 1.13$).

Concerning *like* scores, a Kruskal-Wallis rank test showed a very near to significance difference according to *Strategy* ($H(5) = 10.99$, $p - value = 0.05$). In particular, Dunn's test for multiple comparisons found that **INGR** scores were significantly higher ($M = 3.75$, $SD = 0.62$) than **INTIM** ($M = 2.62$, $SD = 0.96$; $z = 2.87$, p-adj $= 0.03$). No differences were found between **IMPR** scores and the other conditions. In addition, a statistically significant difference between scores was found according to *Nars* scores ($U = 970$, $p = 0.003$): participants who got high scores in the *Nars* questionnaire liked Alice more ($M = 3.62$, $SD = 0.91$) than people who got low scores in the *Nars* questionnaire ($M = 2.92$, $SD = 0.99$).

Concerning *learnfrom, expo* and *rship* (see Table 8.4), no significant differences in scores were found according to any variable. Participants' scores about *learnfrom* and *expo* were all over the mean value, while for *rship* the mean scores for each condition were quite low (all means ≤ 2.75), suggesting that participants considered Alice as very distant from them.

Concerning *likeperson* scores, no significant differences were found according to *Strategy*. Mean scores for each condition were quite low (all means ≤ 2.25), suggesting that in

general Alice was perceived more similar to a computer than a person. A Mann-Whitney U Test showed a statistically significant difference according to *Nars* scores ($U = 1028$, $p = 0.0003$): participants who got high scores in the *Nars* questionnaire perceived Alice less closed to a computer ($M = 2.49$, $SD = 1.12$) than people who got low scores in the *Nars* questionnaire ($M = 1.58$, $SD = 0.69$).

On the whole, these results did not allow us to validate **H2a**, but agent's adaptation was found to have at least an effect on its level of warmth (**H2b**, see subsection 8.5.6.1).

8.5.6.4 Verbal cues of engagement

Only one person gave a negative answer to Topic1_question, while people gave different responses to Topic2_question. We divided participants in two groups, according to the number of verbal feedback they gave to the agent (i.e., if they reply to the agent's speaking turn). In general, participants which did less than 13 verbal feedback (13 was the half of the minimum number of possible speaking turns) out of the 25/36 total possible speaking turns (see Section 8.5.5) gave a positive answer to Topic2_question ($OR = 4.27$, $p = 0.04$). In addition, we found that ratings about *likeperson* item were significantly lower for people giving much verbal feedback ($M = 1$, $SD = 0$) compared to those of people who did not talk a lot ($M = 2.16$, $SD = 1.07$; $U = 36.5$, $p = 0.02$). No differences in any of the dependent variables were found according to *Strategy*.

8.5.7 Discussion

In this section we discuss into details the results of the user study. First of all, regarding **H1**, the only statistically significant results concerned the perception of agent's warmth. Alice was rated as colder when she adopted **INTIM** strategy, compared to the other conditions. This supports the thesis of the primacy of warmth dimension (see Section 3.3), and it is in line with the positive-negative asymmetry effect. In our case, when the agent displayed cold (i.e., low warmth) behaviours (i.e., in **INTIM** condition), it was judged by participants with statistically significant lower ratings of warmth. Regarding the other conditions (**INGR**,**SUPP**,**SELF**, **IMPR** and **RAND**), they elicited warmer impressions in the user, but there was not one strategy better than the others in this regard. The fact that also the **SELF** condition elicited the same level of warmth than the others could reflect an halo effect (see Section 3.4): the behaviours displayed to appear competent influenced its warmth perception in the same direction.

Regarding **H2**, the results did not validate our hypothesis **H2a** that the interaction would be improved when the virtual agent managed its impressions by adapting its strategy according to user's engagement. During the interaction, participants did not show many non-verbal behaviours. Many of them stared at the ECA without moving much. They did not vary their facial expression, move their head or gesture. Since our engage-

ment detection model relied on the interpretation of non-verbal behaviours, the lack of behavioural change impacted directly the output values it returns.

When analysing scores for *perception* items, we found that participants were more satisfied by the interaction and they liked Alice more when it wanted to be perceived as warm (i.e., in **INGR** condition), compared to when it wanted to be perceived cold and competent (i.e., in **INTIM** condition). An idea is that since the agent was perceived warmer in **INGR** condition, it could have positively influenced the ratings of the other items, like *satisfaction*. Concerning **H2b** about a possible effect of agent's adaptation on user's perception of its W&C, it was interesting to see that when the agent adapted its self-presentational strategy according to user's overall engagement, it was perceived as warm. This highlighted a link between agent's adaptation, user's engagement and warm impressions: the more the agent adapted its behaviours, the more the user was engaged and the more s/he perceived the agent as warm.

When looking at participants' verbal cues of engagement (see subsection 8.5.6.4), we could divide people into two groups: those who gave much verbal feedback during the speaking turns, and those who just answered to agent's questions and did not provide verbal feedback during the rest of the interaction. Participants talking a lot may ask questions to the agent, give their opinion on a game, etc. Since the agent was not endowed with natural language understanding capacities, it could not answer participant's request, nor could it argument on user's opinion. Even though we did not explain agent's limitation to participants before starting the experiment, users who gave many feedback at the beginning of the interaction often became aware that the agent could not react to their speech, since it did not consider what they said, interrupt them, continue talking on its topic as if the participants had not talked. This could have had a negative effect on their experience and had led them to choose not to continue to discuss with the agent. When looking at the interaction with this group of people, we noticed that they stopped providing feedback after the virtual agent missed answering them properly. There was a clear distinction in their verbal behaviours before and after the agent missed their input. In our quantitative analyses we found that the majority of people replying a lot to the agent often gave a negative answer to the question Topic2_question asked by the agent about continuing the discussions. On the other hand, people who did not talk a lot had less probability to experience weird situations such as asking a question to the agent and not being heard. These people were less disappointed than the others and more likely to accept to continue the interaction. Indeed, according to our results, the majority of people who did not give much verbal feedback gave a positive answer to the question Topic2_question. This idea that participants giving much feedback at the beginning of the interaction discovered the limits of the agent seemed in line with the lower scores found for *likeperson* item given by people talking a lot compared to the others. The fact that the agent did not behave in the appropriate way and that the agent did not stand up to their expectancies could have highlighted even more the fact that they were in front of a system that simulated a

"mock" of interaction. Another possible explanation to this result could concern the fact that people who did not talk a lot were intimidated and so they did not dare to give a negative answer to the agent. This could be in line too with the results about *likeperson* item: considering the agent closer to a person, they could have answered "yes" as not to offend, somehow, the agent.

In this discussion we should take into account how participants' expectancies may affect their perception of the interaction. People expectancies about others' behaviours have already been demonstrated to affect human-human interaction as well as when people are in front of an ECA (see Section 6.2). In this study we found some effects of people's a-priori about virtual characters: people who got higher scores in the *Nars* questionnaire generally perceived the agent warmer, compared to people who got lower scores in the *Nars* questionnaire. In addition, it should not be forgotten that the fact of being in a Sciences museum, combined with people exposition to films and TV shows about artificial intelligence could have had a strong impact on participants' expectancies. People could have difficulties in distinguishing between what is shown in science-fiction films and the current state of the technology of interactive ECAs. Thus, people could have exaggerated expectancies about our virtual agent's capabilities. These expectancies, and the related disappointment showed by some participants when interacting with a less sophisticated virtual character, could have become an uncontrollable variable preventing any other effect of the independent variables of our experiment. Nevertheless, it has to be remembered that in this experiment we mainly focused on the non-verbal behaviours rather than on the dialogical dimension, limiting therefore the dialogue complexity to better control the other variables. The agent had the floor during the majority of the interaction; our system took into account the polarity of user's answers only at 2 specific moments, Topic1_question and Topic2_question (see subsection 8.5.5), thus the variability of the agent's dialogue was very limited.

Some limitations emerged from the system. First of all, many participants did not like the virtual character, as we could see from their answers to the questionnaires, as well as from their comments during the debriefing. They reported their disappointment about the quality of the animation and of the voice of the agent. They described the experience as "disturbing", "creepy". This could be due to the voice synthesizer and the procedural animation of the agent, that affected the naturalness of agent's behaviour. So probably their very first impression about the appearance and the voice of the agent was too strong and affected the rest of the experience. During the interaction, participants did not show many non-verbal behaviours. This could be due to the setup of the experiment, where participants stood in front of the screen and the virtual agent was displayed at human size. According to their comments, many people were a bit frightened by the dimension of the agent and for almost all of them it was their first interaction with an ECA. We will discuss some possible improvements of the agent's animation and voice in Section 10.3.

8.6 Conclusion

In this Chapter, we presented a computational model for an Embodied Computational Agent, aimed at managing in real-time its self-presentational intentions eliciting different impressions of warmth and competence, in order to maximise user's engagement during the interaction. We built an architecture which took as input participants facial Action Units, trunk and head rotation, used them to compute user's overall engagement and sent it to the Impressions Management Module the agent. Through a multi-armed bandit algorithm which took user's engagement as reward, the agent could select the self-presentational intention maximising user's engagement. In order to evaluate the system, we conceived an interaction scenario where the agent played a role of museum guide. In the experiment we manipulated how the the agent selected its self-presentational intention at each speaking turn. It could adapt its behaviour by using the reinforcement learning algorithm, or choose it randomly, or use the same self-presentational intention during the whole interaction. The agent which adapted its behaviour to maximise user's engagement was perceived as warmer by participants, but we did not find any effect of agent's adaptation on users' evaluation of the interaction.

The key points of this Chapter:

- We customised the general architecture for agent's impressions management in real-time in order to adapt user's self-presentational intentions to mazimise user's engagement.
- We conducted an evaluation study to compare the effects of an adaptive agent and a non-adaptive one on user's impressions of agent's W&C and perception of the interaction.
- We could manipulate users' impression about agent's warmth with different self-presentational techniques.
- We found a link between agent's adaptation, user's engagement and warmth impressions: the more the agent adapted its behaviours, the more the user was engaged and the more s/he perceived the agent as warm.
- People's expectancies about virtual agents affected their answers.

Chapter 9

User Study 2: Adapting agent's behaviour according to user's impressions

Contents

THIS Chapter presents our second use case where we applied the system architecture described in the Chapter 7. We were interested in investigating whether it is possible to affect user's perception of the agent and of the interaction by

adapting the agent's behaviour according to the detected user's impressions. We adapted the modules of the architecture in order to compute user's impressions of agent's warmth and competence from the analysis of their facial expressions. An evaluation study is presented, where we applied our modified architecture in a real-time scenario similar to the one used in the previous study. We compared the effects of an adaptive agent and a non-adaptive one on user's impressions and perception of the interaction.

9.1 Introduction

In the previous Chapter we focused our investigation on the role of warmth and competence in affecting user's engagement during the interaction. In this Chapter we present our second application of the system architecture for ECA's impressions management in real-time. By using the same scenario of the previous use case (with some little changes in the set up, as described in subsection 9.4.5) we aimed to test a detection model which was developed by Chen Wang, Guillaume Chanel and Thierry Pun of the University of Geneva, the partners of the IMPRESSION Project. This model allowed to detect user's impression of agent's warmth or competence by analysing the activity of user's AUs. We have to remember that the self-presentational techniques conveying different levels of W&C were not completely validated in the previous experiment (no significant differences among the different strategies were found for competence ratings and only one technique differed from the others in terms of warmth scores). By exploiting the detection model developed by our partners, our aim was that the agent could manage its behaviour to display the most appropriate impression of warmth or competence.

With this second experiment we aimed to answer to the following research questions:

- (Q1) *Is it possible to elicit different impressions of W&C by adapting the agent's behaviour according to the detected user's impressions?*
- (Q2) *Is it possible to influence user's perception of the interaction by maximizing agent's warmth (or competence) during the interaction?*

In order to answer to these questions, we customised the *User's Analysis Module* (see Section 7.3) in order to compute user's impressions from low-level signals and use them as reward for the RL algorithm. The *Impressions Management Module* (see Section 7.4) was also modified by adapting the reinforcement learning algorithm and including a set of possible verbal and non-verbal behaviours to perform. Differently from the *Self-presentational Intention Instantiation* module used in the previous study (see subsection 8.4.2), in this case we did not create self-presentational intentions but we gave to the agent a set of behaviours that it could combine as it wanted (i.e., by learning from the RL algorithm).

This Chapter is organised as follows: in the next Section we provide a short state of the art about automatic assessment of affective dimensions; in Section 9.3 we describe the

modules of the architecture that have been modified to adapt it to our study; in Section 9.4 we present an evaluation study aiming to investigate our research questions, where we compared an agent adapting its level of warmth or competence with a non-adaptive agent.

9.2 Impressions Assessment

To the best of our knowledge, there is no existing research investigating if impressions of W&C can be assessed from the social signals of the person forming the impression. However, studies in affective computing have demonstrated the possibility to infer user's emotions from multi-modal signals (Brady et al., 2016). Since emotions can be induced when forming impressions (Cuddy et al. (2008), see subsection 3.2.2.2), this supports the possibility of assessing users' impressions from their affective expressions. Emotion recognition studies explored a variety of models using machine learning methods. These methods can be grouped in two classes based on whether temporal information is applied or not. The non-temporal models generally require contextual features while temporal models exploit the dynamic information in the model directly. They include methods such as Multiple Layer Percepton (MLP) or Support Vector Machine (SVM) for example. For temporal models, Long Short Term Memory (LSTM) models are currently widely used with several topologies (Gunes and Pantic, 2010; Brady et al., 2016; Chen et al., 2017). When detecting emotions, different modalities may require various lengths of temporal windows to extract features appropriately (Tzirakis et al., 2017). For example, according to (Gunes and Schuller, 2013; Ringeval et al., 2015), visual modality (upper body recordings) changes faster over time than physiological signals such as heart rate, temperature and respiration rate. There are multiple works from both temporal and non-temporal methods, indicating that facial expression measurements generally achieve better performance as compared with other modalities such as speech and physiological signals for affect recognition (Povolny et al., 2016; Brady et al., 2016).

In the context during which the work presented in this Thesis was developed, there was no existing system to detect W&C impressions given by the agent from user's social signals. The Swiss partner of the project worked with us in parallel to develop such a system. Since when trying to form a mental model of what someone else thinks, the facial expression is the most common studied modality (Baron-Cohen, 1996; de Melo et al., 2014), our system used the user's facial expressions to assess the impression elicited by the agent.

9.3 System Architecture

As mentioned at the beginning of this Chapter, since in the previous study described in Chapter 8 we did not completely validated all the self-presentational strategies of the

agent, our goal at this step was to make the agent learn the verbal and non-verbal behaviour to be perceived as warm or competent by using directly user's impressions as reward. To do this, our architecture needed to contain a module for the detection of user's impressions, and a specific set of verbal and non-verbal behaviours from which the agent could choose.

The modified architecture of the system is depicted in Figure 9.1. In the following Section we give more details about the modified modules.

9.3.1 User's Impressions Detection[1]

A trained Multilayer Perceptron Regression (MLP) model was implemented in the *User's Analysis Module* to detect the impressions formed by users' about the ECA. The MLP model was previously trained with a corpus including face video recordings and continuous self-report annotations of W&C. The model was trained with 32 participants (12 hours recording) watching impression stimuli videos from the NoXi database (see Section 5.2). While the participants were watching the videos, their facial expressions were recorded using a camera (logitech C525 & C920 with sample rate at 30fps) and they were requested to annotate their impressions by pressing buttons when they felt a change in warmth (up & down keyboard arrow) or in competence (left & right keyboard arrow). W&C were annotated independently.

The MLP model had 2 hidden layers and 1 output layer with 50 epochs. A validation set was created with 20% of the training data, to apply early stopping (patience of 5 epochs) and avoid over-fitting. The performance was tested using a leave-one participant out cross-validation which is widely used for small dataset and evaluated using the Concordance Correlation Coefficient (CCC). The average CCC of the MLP model on warmth and competence were 0.64 and 0.70 respectively.

In the *User's Analysis Module*, EyesWeb communicated with the trained MLP model through a TCP connection. EyesWeb implemented a parallel thread to send and receive data to the server. At each Kinect video frame, EyesWeb called OpenFace to get the user's facial AUs configuration. Impression was detected by the MLP model every second with AUs extracted from 30-frame buffer.

More details about the MLP model can be found in (Wang et al., tted).

9.3.2 Impressions Management Module

We modified the reinforcement learning algorithm and the input of the *Communicative Intention Instantiation* module, by giving a set of possible verbal and non-verbal behaviours to perform.

[1]This work has been realized by Chen Wang from University of Geneva.

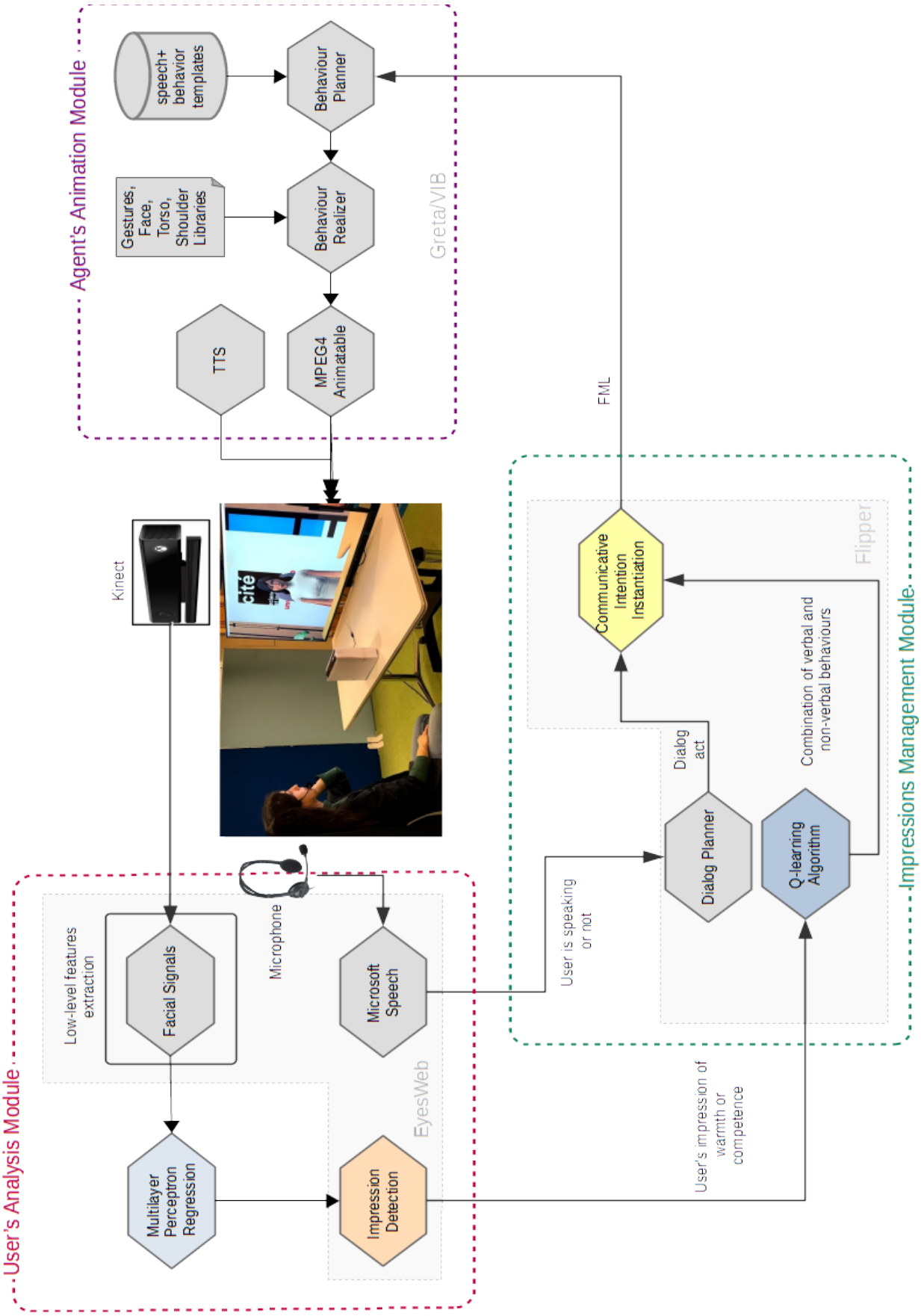


Figure 9.1 – The modified system architecture used in this use case. The modules that have been modified are coloured, while the modules that did not change are grey. In particular, the *User's Analysis Module* contains the model to detect user's impressions from facial signals. The *Impressions Management Module* contains the modified Q-learning algorithm.

9.3.2.1 Reinforcement Learning

To be able to change the ECA behaviour according to detected participant's impressions, we applied a Q-learning algorithm. In our case, states *s* were the agent's warmth/competence level and the actions *a* performed by the agent concerned the verbal and non-verbal behaviours listed in subsection 9.3.2.2. The initial Q values $Q(s, a)$ of actions and states were set up to 0. The reward function R computed for each combination of state and action was the difference between detected warmth (resp. competence) and the current warmth (resp. competence) level. The Q-learning algorithm explored all the possible next state-action pairs s', a' and tried to maximize the future rewards with a discount rate γ. We maximized one dimension at a time since the MLP model gave us the score about only one dimension. The new Q values $Q_{\text{new}}(s, a)$ were updated with the Q function:

$$Q_{\text{new}}(s, a) = Q(s, a) + \alpha[R(s, a) + \gamma \max_{a} Q'(s', a') - Q(s, a)] \tag{9.1}$$

where α was the learning rate, and $Q(s, a)$ was the Q value of current state and action.

9.3.2.2 Communicative Intention Instantiation

The combination of verbal and non-verbal behaviours selected by the Q-learning algorithm and sent to the Intention Planner concerned the same variables manipulated in the previous study described in Chapter 8:

- *Type of gestures*. The ECA could perform *ideational* or *beat* gestures or *no gestures*.
- *Arms rest poses*: in the absence of any kind of gesture, these rest poses could be performed by the ECA: *akimbo* (hands on the hips), *arms crossed* on the chest, *arms along* its body, or *hands crossed* on the table.
- *Smiling*. During the animation, the ECA could decide whether or not to perform smiling behaviour, characterized by the activation of AU6 and AU12.
- *Verbal behaviour*. We used the 4 different verbal behaviours used in the previous study (see Section 8.3), characterised by the different use of *you*- and *we*-words, the level of formality of the language, the length of the sentences.

In this case we did not create combinations of these variables in order to create self-presentational strategies but the agent could perform every possible combination of behaviours. We let the agent learn the best combinations according to its goal to be perceived as warm or competent.

9.4 User Study

We conducted a user study in order to test our model in a user-agent real-time interaction scenario. The aim of the study was to investigate whether the adaptation of the agent's verbal and non-verbal behaviour through our Q-learning algorithm could impact user's impressions of the ECA's W&C and user's overall perception of the interaction.

9.4.1 Independent Variables

The independent variable manipulated in this study concerned the use of our Q-learning model (*Model*), and included 3 conditions:

- *Warmth*, when the agent adapted its behaviours according to user's warmth impressions, with the goal to maximise its warmth;
- *Competence*, when the agent adapted its behaviours according to user's competence impressions, with the goal to maximise its competence;
- *Random*, when the model was not exploited and the agent randomly chose its behaviour, without considering user's reactions.

9.4.2 NARS

Before the interaction with the ECA, we asked participants to fill in the adapted version of NARS scale from Nomura et al. (2006) that was used in the previous study to assess their a-priori about virtual characters (*NARS*).

9.4.3 Dependent Variables

After the interaction with the ECA, participants were asked to fill in some questionnaires where they rated the agent's W&C and their overall satisfaction of the interaction. The dependent variables measured during the study were:

- User's perception of agent's warmth (*w*) and competence (*c*): participants were asked to rate their level of agreement about how well each adjective described the agent (the same used in the previous study, 4 concerning warmth, 4 concerning competence, according to Aragonés et al. (2015)).
- User's perception of the interaction (*perception*): participants were asked to rate their level of agreement about the same list of items described in subsection 8.5.3.2 (*satisfaction*, *continue*, *like*, *learnfrom*, *expo*, *rship*, *likeperson*).

9.4.4 Hypotheses

We hypothesised that:

- **H1**: The ECA would be perceived *warmer* when it adapted its behaviours according to user's warmth impressions, that is, in the *Warmth* condition, compared to the *Random* condition;
- **H2**: The ECA would be perceived *more competent* when it adapted its behaviours according to user's competence impressions, that is, in the *Competence* condition, compared to the *Random* condition;
- **H3**: When the ECA adapted its behaviours, that is in either *Warmth* and *Competence* conditions, this would improve user's overall experience, compared to the *Random* condition.

9.4.5 Procedure

We used the same scenario of the previous study, where the agent played the role of the virtual museum guide Alice giving information about the video games exhibit. We slightly improved the setup by taking into account some comments of the participants of the previous experiment. In particular, we changed the position of the agent from standing to sitting at the desk, with the purpose of putting the user more at ease. We also changed the hair colour of Alice from blond to brown, since many participants of the previous experiment reported that they associated the blond character to the stereotype of a stupid girl. The final adjustment we applied concerned the dialogue of the agent: we reduced the number of steps to 26, as depicted in Figure 9.2. We introduced the step *Discussion*, where the agent asked to the users to describe their favourite game, in order to make they feel more involved in the dialogue. In addition, the agent asked the *Topic2_question* relative to the willingness of the user to continue to discuss about the exhibit. The polarity of user's answer did not affect the following reply of the agent, that could fit both user's positive and negative answer. For example, the agent could answer "I am very talkative, I want to tell you everything" or "Whatever you say, I want to tell you more". We made this choice in order to prevent any detection error that could negatively affect the interaction.

The study took around 15 minutes and was conducted as follows:

1. At the beginning, the participant sat at the questionnaires' place, read and signed the consent form, and filled the *NARS* questionnaire [5 min];
2. The participant then moved to the center of the room, and sat in front of a desk and a big screen displaying Alice. The agent was sitting at a virtual desk placed at the same level than the participant. At the top of the screen, a Kinect 2 was installed, as depicted in Figure 9.3. At the other side of the desk, a black tent was

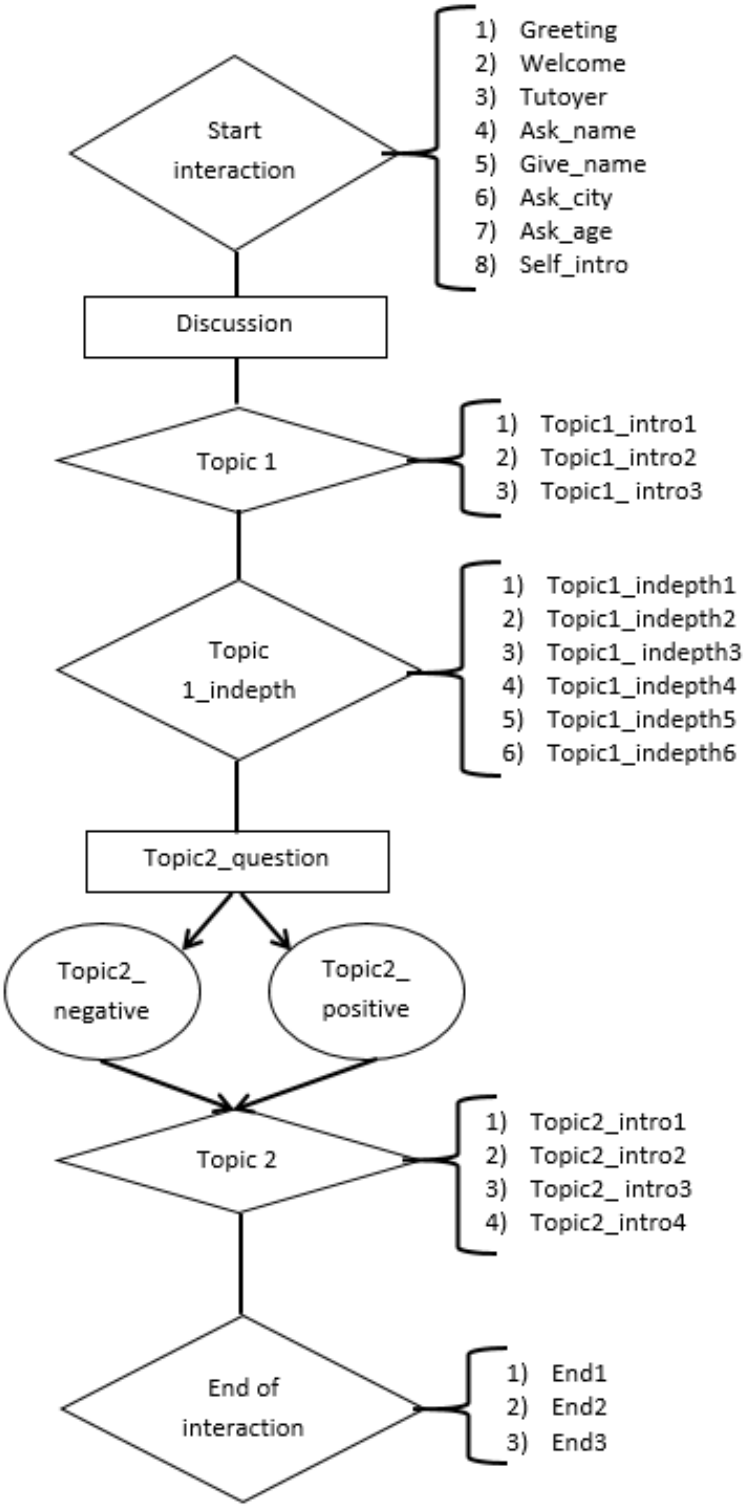


Figure 9.2 – The dialogue flowchart. The diamond shapes represent the main parts that contain several steps. The rectangles represent questions and the rounds represent user's reply to agent's question.

installed, in order to help the Kinect's detection of the user. During the interaction, the participant wore a headset and was free to interact with the ECA as she wanted. The experimenter stayed in a hidden place behind the screen [3 min];

3. The last step consisted in filling in the last questionnaires and debriefing the participant [5-7 min].

Figure 9.3 – The set up of the study: in the foreground, the desk and the screen where the interaction took place; in the background, the questionnaire place with a laptop used to answer to the questionnaires.

9.4.6 Analysis and Results

We collected data from 71 participants, 34% of them were women. Participants were visitors of the Carrefour Numérique of the Cité des sciences et de l'industrie, and were invited to take part to a research study. 28% of them were in the range of 18-25 years old, 18% were in the range 25-36, 28% in the range 36-45, 15% in the range of 46-55 and 11% over 55 years old. Participants were randomly assigned to each condition. In total, 25 participants were assigned to the *Warmth* model, 27 to the *Competence* model and 19 to the *Random* one.

In order to group together the 4 items for *w* and the 4 for *c*, we computed Cronbach's alphas on their scores: good reliability was found for both ($\alpha = 0.85$ and $\alpha = 0.81$ respectively). Then we computed the mean of these items in order to have one *w* score and one *c* score for each participant and used them for our analyses.

Since *NARS* scores got an acceptable score of reliability ($\alpha = 0.69$), we computed the overall mean of these items for each participant and divided them into 2 groups, "high" and "low", according to whether they obtained a score higher than the overall mean or not, respectively. Participants were almost equally distributed into the two groups (35 in the "high" group, 36 in the "low" group). Chi-square tests for *Model*, age and sex were run to verify that participants were equally distributed across these variables, too (all $p > 0.5$).

9.4.6.1 Warmth's Scores

Since *w* means were normally distributed (Shapiro test's $p = 0.07$) and their variances homogeneous (Bartlett tests' ps for each variable were > 0.44), we run 3x5x2x2 between-subjects ANOVA, with *Model*, age, sex and *NARS* as factors.

No effects of age or sex were found. A main effect of *NARS* was found ($F(1, 32) = 4.23, p < 0.05$). Post-hoc test specified that the group who got high scores in *NARS* gave higher ratings about Alice's *w* ($M = 3.65, SD = 0.84$) than the group who got low scores in *NARS* ($M = 3.24, SD = 0.96$).

Although we did not find any significant effect, *w* scores were on average higher in *Warmth* and *Competence* conditions than in the *Random* condition. Mean and standard error of *w* scores are shown in Table 9.1 and Figure 9.4.

Model	**Warmth** $\mu \pm SD$	**Competence** $\mu \pm SD$
WARMTH	3.48 ± 0.8	3.2 ± 0.75
COMPETENCE	3.51 ± 0.96	3.3 ± 0.69
RANDOM	3.26 ± 0.93	2.761 ± 0.73

Table 9.1 – Mean and standard deviation of warmth and competence scores for each level of *Model*.

9.4.6.2 Competence's Scores

Since *c* means were normally distributed (Shapiro test's $p = 0.22$) and their variances homogeneous (Bartlett tests' ps for each variable were > 0.25), we run 3x5x2x2 between-subjects ANOVA, with *Model*, age, sex and *NARS* scores as factors.

We did not find any effect of age, sex or *NARS*. A strong tendency towards statistical significance was found for a main effect of the *Model* ($F(2, 32) = 3.22, p = 0.0471, \eta^2 = 0.085$). In particular, as shown in Figure 9.4, post-hoc tests revealed that participants in the *Competence* condition gave higher scores about Alice's *c* than participants in the *Random* condition ($M1 = 3.3, M2 = 2.76, p\text{-}adj = 0.05$).

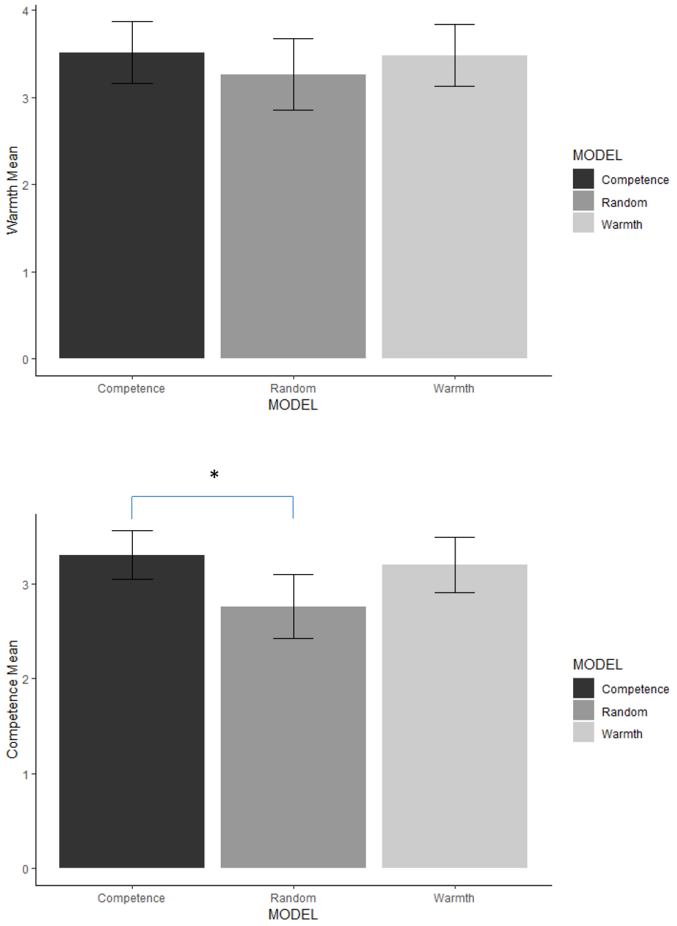


Figure 9.4 – Warmth and competence means for each level of *Model*. * stands for $p = 0.05$.

9.4.6.3 Perception Scores

Since *perception* items' means were not normally distributed but their variances were homogeneous (Bartlett tests' ps for each variable were > 0.17), we run non-parametric tests for each item and each variable.

Even if we did not find any statistically significant effect, on average items' scores tended to be higher in *Warmth* and *Competence* conditions than in *Random* condition.

9.4.7 Discussion

The results showed that participants' ratings tended to be higher in the conditions in which the agent used the Q-algorithm to adapt its behaviour, compared to when it selected its behaviour randomly. In particular, the results indicated that we successfully manipulated the impression of competence when using our adaptive ECA. Indeed, higher competence was reported in the competence condition compared to the random condition. No a-priori effect for competence was found.

On the other hand, we found an a-priori effect on warmth but no significant effect of our conditions (just a positive trend for both competence and warmth conditions). People with high a-priori about ECAs gave higher ratings about Alice's warmth than people with low a-priori.

We could hypothesise some explanations about these results. First, we did not get effects of our experimental conditions on warmth ratings since people were more anchored into their a-priori and it was hard to change them. This is in line with literature (Burgoon et al., 2016) and our previous results of our perceptive study described in Chapter 6. The fact that we found this effect only for warmth judgments could be related to the primacy of warmth judgments over competence (see Section 3.3). Then, it could have been easier to elicit impressions of competence since we found no a-priori effect on competence. This could be explained as people might expect that it is easier to implement knowledge in an ECA rather than social behaviours.

Concerning users' perception of the interaction, we tried to understand why it only tended to be different across conditions but did not reach statistical significance. During the debriefing, as in the previous experiment, many participants told us their disappointment about Alice's appearance, the quality of the voice synthesizer and the animation, as well as the limitations of the conversation (participants could only answer to Alice's questions). This deception could have reduced any other effect of the independent variables. Indeed, Alice's appearance and the structure of the dialogue were the same across conditions. If participants mainly focused on these elements, they could have paid less attention to ECA's verbal and non-verbal behaviour (the variables that were manipulated and we were interested in), which thus did not manage to affect their overall perception of the interaction.

9.5 Conclusion

In this Chapter we presented the second application of the system architecture for agent's impression management. We exploited a model to detect user's impressions about agent's warmth or competence during the interaction and we used them as the reward of a Q-learning algorithm in the Impressions Management Module. In this use case we did not create self-presentational strategies conveying different levels of warmth and competence. Instead, the agent could learn how to combine together its different verbal and non-verbal behaviours, according to the detected user's impressions. We conducted an evaluation study where we compared an agent adapting its level of warmth or competence with a non-adaptive agent. Results showed that the adaptive agent managed to influence user's impressions about its competence, while users' a-priori affected their impressions about agent's warmth.

The key points of this Chapter:

- We customised the general architecture for agent's impressions management in real-time in order to maximise user's impressions of agent's warmth and competence by detecting user's impressions about the agent from their facial signals.
- We conducted an evaluation study to compare the effects of an adaptive agent and a non-adaptive one on user's impressions of agent's W&C and perception of the interaction.
- Participants' ratings tended to be higher in the conditions in which the agent adapted its behaviour, compared to when it selected its behaviour randomly.
- We successfully manipulated the impression of competence when using our adaptive ECA.
- People's a-priori about virtual agents affected their perception of agent's warmth.

Part VII

Conclusion

Chapter 10

Conclusion and Perspectives

> There are two possible outcomes: if the result confirms the hypothesis, then you've made a measurement. If the result is contrary to the hypothesis, then you've made a discovery.
>
> Enrico Fermi

Contents

In the work presented in this Thesis we contributed to the implementation of a system to endow an Embodied Conversational Agent with the capability to adapt its impressions of warmth and competence according to user's reactions in real-time. The architecture of the system includes an Impressions Management Module which uses user's reactions, such as for example her perceived engagement level or impressions of agent's warmth and competence, as a reward for the selection of the best verbal and non-verbal behaviours characterising the agent's dialog act. To select the set of possible

behaviours that serves as basis for agent's learning, we followed an approach that started from the analysis of human-human interactions in order to investigate the role of non-verbal behaviours such as gestures, rest poses, smiling, head movements in impressions formation. We finally evaluated this system in two real scenarios where the agent played the role of virtual guide of a museum adapting its self-presentational strategies in order to maximise user's behaviour, or adapting its verbal and non-verbal behaviour in order to maximise the impressions of warmth and competence elicited in the user.

In this Chapter we conclude this Thesis by summarising our contributions in Section 10.1, by identifying some limitations of this work in Section 10.2 and by suggesting future perspectives to improve our work in short- and long-term in Section 10.3.

10.1 Summary of Contributions

First contribution: *Creation of a repertoire of multi-modal behaviours eliciting impressions of warmth and competence.*

During the first phase of our work, our purpose was to find associations between non-verbal behaviours associated to warmth and competence impressions. We started from the study of literature about non-verbal cues of warmth and competence and existing studies which included these dimensions in human-agent interaction. After this first investigation, we followed an approach led by 2 main motivations. First, we did not find a great quantity of information from literature about non-verbal behaviour eliciting warmth and competence impressions. Second, the few works concerning non-verbal behaviour of a virtual agent eliciting warmth and competence impressions relied on corpus of actors, while we wanted to collect information from the study of natural interactions. That's why we annotated and analysed the NoXi corpus, as described in Chapter 5.

The annotations added to the corpus are available in https://noxi.aria-agent.eu/ and they can be useful for other researchers for further analyses or to produce more annotations by following the same annotation schema we used (see subsection 5.3.2). We contributed to give insights about the role of type of gestures, arms rest poses and smiling behaviour in forming these impressions. We found some matches with literature such as the presence of halo effect (Rosenberg et al., 1968) for gestures and compensation effect (Yzerbyt et al., 2008) for smiling. Another contribution comes from the results of the perceptual study presented in Chapter 6, which highlighted the role of expectancies when judging virtual agents, in line with the theory of Burgoon et al. (2016).

Second contribution: *Creation of a behaviour adaptation module for ECA's impressions management.*

We created a module based on reinforcement learning allowing the ECA to adapt its behaviour to user's reactions. It allows defining different rewards for the behaviours used by the agent. This module allows learning in real-time without having previous knowledge about user's reactions to its behaviour. It allows testing all the possible combinations of verbal and non-verbal behaviours to find the best one to produce a certain impression on the user. It can be adapted to different goals of the agent and it allows having better insights about the role of non-verbal behaviours in human-agent interaction.

Third contribution: *Creation of a set of strategies for managing impressions of warmth and competence in an ECA.*

Starting from the findings coming from the analysis of human-human interaction, we investigated the role of multi-modal behaviours and expectancies when judging virtual agents, in order to create a set of strategies for managing impressions of warmth and competence in an ECA. These strategies were inspired by the taxonomy of Jones and Pittman (1982). According to the chosen strategy, the agent perform verbal and non-verbal behaviour with the goal to be perceived as warm, competent, warm and not competent, or cold and competent. These strategies were partially validated in our evaluation study, especially for the warmth dimensions, and could be implemented in the general architecture for agent's impressions management.

Fourth contribution: *Implementation of the impressions management module in a system architecture for real-time user-agent interaction.*

We integrated the reinforcement learning module for agent's impressions management in a system architecture. This included a module to detect and interpret user's multi-modal reactions and a module for agent's behaviour generation. The architecture is general enough to allow for customisation of the different modules according to different contexts and goals of the agent. With this work, we gave a strong contribution by conceiving a human-agent interactive framework that can be adapted and exploited in further projects. For example, the possibility to implement modules for detecting user's physiological data would allow obtaining a better understanding of the impact of agent's behaviour on user's affective state. This work offers potential big impact on many applications such as web assistant agents and real life agent installations (e.g. in train stations or museums).

Fifth contribution: *Investigation of the effectiveness of an adaptive agent and of the relationship between agent's adaptation, engagement and impressions, during human-agent interaction.*

We conceived a scenario where the virtual agent played the role of virtual museum guide

and personalised the general architecture for impressions management in 2 different applications. In the first one the agent adapted its self-presentational strategies to be perceived more or less warm or competent with the goal to maximise user's engagement. In the second case it adapted its behaviour in order to maximise user's impressions of its warmth or competence level. We designed and conducted an evaluation study for each scenario in order to validate the effectiveness of the agent's impression management capabilities. In particular we wanted to verify that users preferred an agent with impressions management skills over an agent which did not manage users' impressions. The experimental studies conducted at the Cité des sciences et de l'industrie were crucial to understand what people really expected from Embodied Conversational Agents and what it should be improved in our model. We will discuss these points in the following Section.

Some interesting results emerged from our studies, showing some effectiveness of the model. In particular, we found that a link between agent's adaptation, user's engagement and warm impressions: in the study presented in Chapter 8 the more the agent adapted its behaviours, the more the user was engaged and the more s/he perceived the agent as warm. In the study described in Chapter 9 we also found a tendency for participants to give higher ratings when the agent used the reinforcement learning algorithm to adapt its behaviour, compared to when it selected its behaviour randomly. In particular, the results indicate that we successfully manipulated the impressions of competence when using our adaptive agent. On the other hand, we found an a-priori effect on warmth but no significant effect of our conditions (just a positive trend). People with high a-priori about ECAs gave higher ratings about Alice's warmth than people with low a-priori.

10.2 Limitations of our work

The contributions presented in this Thesis are not exempted from limitations. We identified some characteristics of our methodology and our approach that could be improved. In the following paragraphs we discuss some limitations concerning the annotation methodology (subsection 10.2.1), the use of the reinforcement learning algorithm for managing agent's impressions (subsection 10.2.2) and the scenarios used in our experimental studies (subsection 10.2.4).

10.2.1 Corpus Annotation

We opted for manual annotation of the NoXi corpus instead of exploiting automatic tools, since the error rate of these tools was quite high to allow for reliable annotations, especially for arms rest poses and head movements. Our choice was well motivated but at the same time it required a huge amount of time to obtain a limited quantity of data available for the analysis.

Another limitation of our methodology concerns the continuous annotations of perceived warmth and competence level. These annotations are subjective, thus prone to biases caused by, for instance, annotator's tiredness or social desirability. In addition, in order to obtain highly reliable annotations with agreement we needed to drastically reduce the amount of data for our analysis, which might have prevented us to find more statistical relationships between warmth and competence impressions and non-verbal behaviours.

Finally, verbal behaviour was not considered during annotations. Voice prosodic features as well as speech's content could have been collected for looking at the role of both verbal and non-verbal cues in the impression formation of warmth and competence.

10.2.2 Impressions Management Module

The use of reinforcement learning in our Impressions Management Module was motivated by the need to learn in real-time without previous knowledge about the user. However, this algorithm could have some limitations when the set of possible behaviours from which to learn is huge and the number of steps for each learning episode is reduced. In addition, the reinforcement learning algorithm is strongly tied to the user's detection module. If the model to detect user's reaction is not reliable enough, and it would not distinguish different user's states, it would be difficult for the reinforcement learning algorithm to learn which behaviour is better than another one if receiving the same reward for every behaviour using human actors.

10.2.3 Agent's Animation and Voice

As many participants told us during the debriefing, the quality of agent's animation and voice was often perceived as unnatural. This is due to the animation method and the voice synthesizer used in the platform Greta/VIB.

The animation of the agent is realised through a method that combines procedural and key-frame techniques. In key frames animation experts create particular steps of the animation, that are then linked together by interpolation. By combining a library of key frames with procedural animation, the agent can dynamically modify them in real-time, for example through expressivity parameters. This combined method makes the animation more flexible than key frame alone but still not realistic enough compared to motion capture. Participants' comments revealed that this method, although potentially powerful, is not good enough to produce natural animation, especially for people used to video games and virtual reality where characters animation is done by motion capture. Motion capture produces very natural animation but it is very expensive method that needs to create every possible gesture.

The voice synthesizer of the agent did not produce a very realistic voice either. That is due to similar reasons to those concerning the animation method. The tool is powerful

since it generates the agent's sentence dynamically in real-time, that is, we can set several parameters such as emotion or prosody. It learns a voice model which binds phonemes to create words. But still the voice is synthetic and not human. Indeed, is it impossible to record every possible sentence from a human to be used in a real-time interaction system.

10.2.4 Evaluation Studies

The protocol followed in our evaluation studies did not include a natural dialogue between the user and the agent, since it was beyond our research interests. This lack of verbal interaction had an impact on participant's experience, as showed in subsection 8.5.7. In addition, when conceiving the dialog steps, we did not focus on building a rapport with the participant: the agent just managed its impressions of warmth and competence without considering the social relation with the user. Rapport, meant as the feeling of harmony and connection with another, is an important aspect of human interaction, as well as of human-agent interaction (Gratch et al., 2007; Zhao et al., 2016). Agent's communicative intentions should take into account this dimension, at both verbal and non-verbal level. For example, we could include some conversational strategies such as self-disclosure, enhance the gaze behaviour of the agent to improve mutual attentiveness, and provide agent's non-verbal listening feedback, such as postural mimicry and synchronisation of its movements with the user's ones.

10.3 Perspectives

In this section we envisage some improvements to our work in a short- and long-term perspective. Some of them could go beyond the limitations discussed in the previous Section, while other would allow us to apply our model to further investigations.

10.3.1 Short-term

10.3.1.1 Animation and voice improvement

The first improvement that could be done concerns the appearance, the quality of the animation and the voice of the agent. The use of other techniques such as motion capture could improve the quality of the animation. Concerning the agent's voice, since the voice synthesizer that we used limited the naturalness of the voice, we could exploit more realistic tools such as Google Cloud Text-to-Speech [1].

[1] https://cloud.google.com/text-to-speech/

10.3.1.2 Implementation of a Natural Language Processing Module

As mentioned in subsection 10.2.4, we did not implement a complex module for natural language processing. We kept the dialogue at a very simple level since it was not the scope of our work. In our scenario the agent led the conversation and in some cases considered the polarity of the answers of the users but not the content of their speech. We could improve the quality of the interaction by integrating in our architecture a more complex natural language processing tools. For example, we could add a set of topic transition strategies (Glas and Pelachaud, 2015c), or detect user's opinion (Barriere et al., 2018).

10.3.1.3 Behaviour adaptation of backchannels

Whereas our Impressions Management Model focused on behavioural adaptation of the agent while speaking to the user, an important role is played by backchannels (Yngve, 1970), that have been the object of many researchers (see for example Bevacqua et al. (2010b); Poppe et al. (2011)). These could include head movements, smiling, vocalisations like "yeah", "mhmh" or even laughter. They can convey different meanings, such as agreement, refusal, liking, interest, etc.. They influence user's perception, for example smile backchannels make an agent warmer and more positively judged (Bevacqua et al., 2010a). A higher number of generated backchannels increases the naturalness of the backchannel behaviour (Poppe et al., 2013). Agent's backchannels could be integrated to our Impressions Management Module, in order to give the agent the ability to adapt its behaviours also while listening.

10.3.2 Long-term

10.3.2.1 Learning from multiple sources

In a long-term perspective, it would be interesting to detect many high-level features from users at the same time and apply multiple reinforcement learning techniques (Shelton, 2001) in the Impressions Management Module. For example, we could integrate together the engagement detection model used in Chapter 8 and the impressions detection model used in Chapter 9. This would allow to detect at the same time user's engagement level and its impressions about the agent. We could obtain more insights about the relationship between engagement and warmth and competence impressions than those that we already found in our experimental study described in Chapter 8.

10.3.2.2 Use of more than one agent during the interaction

Another future direction of our work concerns the realisation of a scenario with more than one agent, by exploiting Greta/VIB version for multi-characters animation. Multiple

agents setting has been shown to be more effective than a single agent on user's persuasion (Kantharaju et al., 2018). It would be interesting to investigate whether it can affect user's impressions of the agent, too. For example, we could investigate if we can influence the impressions about one agent by manipulating the behaviours of another agent. In addition, by applying Judd et al. (2005) paradigm to our context, we could make one agent learn to maximise its impressions about one dimension in one direction, and make the other agent learn to maximise it in the opposite direction, and investigate if we can induce a compensation effect on user's judgements about the non-manipulated dimension.

10.3.2.3 Investigating the dynamics of first impressions

In the work presented in this Thesis we analysed impressions over a time-window of a few minutes: we analysed the first 5 minutes of the NoXi database, we conceived experimental scenarios lasting 3 minutes. Once we assessed user's impressions about the agent, it would be interesting to investigate the dynamics of these impressions during time. For example, how long this impression lasts, how difficult it is to change it during a second interaction. Bergmann et al. (2012) found that warmth judgements are more flexible than competence ones. In their experiment (see Section 4.1) they asked participants to judge the virtual agent at two different points of measurements (one after 15 seconds and one after few minutes). They found that users' judgements about warmth decreased from the first to the second measuring point for robot-like agent, while human-like agents provided more stable impressions of warmth. This did not occur for competence ratings. They also found that use of gestures increased competence ratings between measurement points, while the absence of gestures resulted in a decrease of competence ratings. It seems thus that impressions could be modified over time. While the measurement performed by Bergmann et al. (2012) concerned a brief interval of time, it would be interesting to measure users' impressions also after a long period, such as weeks, and compare impressions over time of an adaptive agent with a non-adaptive one.

Part VIII

Annexes

Appendix A

Publications and Dissemination

The publications in International Conferences and Workshops include:

- Biancardi, B., Wang, C., Mancini, M., Cafaro, A., Chanel, G., and Pelachaud, C. (2019). A computational model for managing impressions of an embodied conversational agent in real-time. In *2019 International Conference on Affective Computing and Intelligent Interaction (ACII)*
- Mancini, M., Biancardi, B., Dermouche, S., Lerner, P., and Pelachaud, C. (2019). An architecture for agent's impression management based on user's engagement. In *International Conference on Intelligent Virtual Agents*. Springer
- Biancardi, B., Cafaro, A., and Pelachaud, C. (2017a). Analyzing first impressions of warmth and competence from observable nonverbal cues in expert-novice interactions. In *19th ACM International Conference on Multimodal Interaction*. ACM
- Biancardi, B., Cafaro, A., and Pelachaud, C. (2017b). Could a virtual agent be warm and competent? investigating user's impressions of agent's non-verbal behaviours. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*, pages 22–24. ACM
- Beatrice, B., Cafaro, A., and Pelachaud, C. (2017). Investigating the role of gestures, arms rest poses and smiling in first impressions of competence. In *3rd WS on Virtual Social Interaction*

The publications in National Workshops include:

- Biancardi, B., Cafaro, A., and Pelachaud, C. (2018). Étude des effets de différents types de comportements non-verbaux sur la perception d'un agent virtuel. *Workshop Affect, Compagnon Artificiel, Interaction (WACAI)* **BEST PRESENTATION AWARD**

- Biancardi, B., Cafaro, A., and Pelachaud, C. (2016). Investigating user's first impressions of a virtual agent's warmth and competence traits. In *Workshop Affect, Compagnon Artificiel, Interaction (WACAI)*

The publications in International and National Doctoral Consortia include:

- Biancardi, B. (2017). Towards a computational model for first impressions generation. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 628–632. ACM
- Biancardi, B., Cafaro, A., and Pelachaud, C. (2017c). Gérer les premières impressions de compétence et de chaleur à travers des indices non verbaux. In *Quatorzièmes Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA 2017)*

The publications submitted to an International Journal and an International Conference include:

- Biancardi, B., Mancini, M., Lerner, P., and Pelachaud, C. (under revision). Managing an agent's self-presentational strategies during an interaction. *Frontiers in Robotics and AI - Computational Approaches for Human-Human and Human-Robot Social Interactions*
- Wang, C., Chanel, G., Biancardi, B., Mancini, M., and Pelachaud, C. (submitted). Automatic impression detection and a use case with an embodied conversational agent. In *21th ACM International Conference on Multimodal Interaction*. ACM

In July 2018 I was invited to give the talk "Vers des Agents Conversationnels Animés plus chaleureux et compétents" in the context of the Symposium "Ils sont peu compétents mais si chaleureux! Pourquoi et quand le phenomène de compensation guide nos jugements sociaux" of the 12e Congrès International de Psychologie Sociale en Langue Française, in Louvain-la-Neuve, Belgium.

In June 2018 I was invited to present my PhD work during the Atelier "Explorer les interactions sociales conversationnelles avec des agents artificiels" - JEP 2018, in Aix-en-Provence, France.

The work presented in this Thesis has been presented to non-expert audience in many occasions, including:

- The 2nd day of the Italian research in the world, at the Italian Embassy in Paris, where I won the 3rd prize for poster presentation;
- An interview during the radio show "La Méthode Scientifique" of France Culture;
- A debate about the future of our relationships with artificial companions during the "Fete de la Science";

- The participation in the competition "Ma thèse en 180 secondes", where I won the 1st Prize of the Jury at Sorbonne University;
- The dissemination of my research while running the experimental studies at the Cité des sciences et de l'industrie in Paris.

Bibliography

Abele, A. E. and Bruckmüller, S. (2011). The bigger one of the "big two"? preferential processing of communal information. *Journal of Experimental Social Psychology*, 47(5):935–948.

Abele, A. E. and Wojciszke, B. (2013). The big two in social judgment and behavior. *Social Psychology*, 44(2):61–62.

Abele, A. E. and Wojciszke, B. (2014). Communal and agentic content in social cognition: A dual perspective model. In *Advances in experimental social psychology*, volume 50, pages 195–255. Elsevier.

Acosta, J. C. and Ward, N. G. (2011). Achieving rapport with turn-by-turn, user-responsive emotional coloring. *Speech Communication*, 53:1137–1148.

Afifi, W. A. and Burgoon, J. K. (2000). The impact of violations on uncertainty and the consequences for attractiveness. *Human Communication Research*, 26(2):203–233.

Aggarwal, P., Artstein, R., Gerten, J., Katsamanis, A., Narayanan, S., Nazarian, A., and Traum, D. R. (2012). The twins corpus of museum visitor questions. In *LREC*.

Allwood, J., Nivre, J., and Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *J. Semantics*, 9:1–26.

Ambady, N. and Skowronski, J. J. (2008). *First impressions*. Guilford Press.

Anderson, N. H. (1962). Application of an additive model to impression formation. *Science*, 138(3542):817–818.

Anderson, N. H. (1968). Application of a linear-serial model to a personality-impression task using serial presentation. *Journal of Personality and Social Psychology*, 10(4):354.

Anderson, N. H. and Alexander, G. R. (1971). Choice test of the averaging hypothesis for information integration. *Cognitive Psychology*, 2(3):313–324.

Andre, E., Rehm, M., Minker, W., and Bühler, D. (2004). Endowing spoken language dialogue systems with emotional intelligence. In *Tutorial and Research Workshop on Affective Dialogue Systems*, pages 178–187. Springer.

Aragonés, J. I., Poggio, L., Sevillano, V., Pérez-López, R., and Sánchez-Bernardos, M.-L. (2015). Measuring warmth and competence at inter-group, interpersonal and individual levels/medición de la cordialidad y la competencia en los niveles intergrupal, interindividual e individual. *Revista de Psicología Social*, 30(3):407–438.

Argyle, M. (1975). *Bodily communication*. Methuen Publishing Company.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3):258.

Aubrey, A. J., Marshall, D., Rosin, P. L., Vendeventer, J., Cunningham, D. W., and Wallraven, C. (2013). Cardiff conversation database (ccdb): A database of natural dyadic conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 277–282.

Bakan, D. (1966). The duality of human existence: An essay on psychology and religion.

Ball, G. and Breese, J. (2000). Relating personality and behavior: posture and gestures. In *Affective interactions*, pages 196–203. Springer.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE.

Baron-Cohen, S. (1996). Reading the mind in the face: A cross-cultural and developmental study. *Visual Cognition*, 3(1):39–60.

Barriere, V., Clavel, C., and Essid, S. (2018). Attitude classification in adjacency pairs of a human-agent interaction with hidden conditional random fields. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4949–4953. IEEE.

Baumgardner, A. and Levy, P. (1987). Interpersonal reactions to social approval and disapproval: The strategic regulation of affect. *Manuscript submitted for publication, Michigan State University*.

Baur, T., Mehlmann, G., Damian, I., Lingenfelser, F., Wagner, J., Lugrin, B., André, E., and Gebhard, P. (2015). Context-aware automated analysis and annotation of social human–agent interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(2):11.

Bayes, M. A. (1972). Behavioral cues of interpersonal warmth. *Journal of Consulting and clinical Psychology*, 39(2):333.

Baylor, A. L. and Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education*, 15(2):95–115.

Beatrice, B., Cafaro, A., and Pelachaud, C. (2017). Investigating the role of gestures, arms rest poses and smiling in first impressions of competence. In *3rd WS on Virtual Social Interaction*.

Beck, R. C. (2003). *Motivation: Theories and principles, 4/e*. Pearson Education India.

Bergmann, K., Eyssel, F., and Kopp, S. (2012). A second chance to make a first impression? how appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. In *International Conference on Intelligent Virtual Agents*, pages 126–138. Springer.

Bevacqua, E., Hyniewska, S. J., and Pelachaud, C. (2010a). Positive influence of smile backchannels in ecas. In *International Workshop on Interacting with ECAs as Virtual Characters*, page 13.

Bevacqua, E., Pammi, S., Hyniewska, S. J., Schröder, M., and Pelachaud, C. (2010b). Multimodal backchannels for embodied conversational agents. In *International Conference on Intelligent Virtual Agents*, pages 194–200. Springer.

Biancardi, B. (2017). Towards a computational model for first impressions generation. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 628–632. ACM.

Biancardi, B., Cafaro, A., and Pelachaud, C. (2016). Investigating user's first impressions of a virtual agent's warmth and competence traits. In *Workshop Affect, Compagnon Artificiel, Interaction (WACAI)*.

Biancardi, B., Cafaro, A., and Pelachaud, C. (2017a). Analyzing first impressions of warmth and competence from observable nonverbal cues in expert-novice interactions. In *19th ACM International Conference on Multimodal Interaction*. ACM.

Biancardi, B., Cafaro, A., and Pelachaud, C. (2017b). Could a virtual agent be warm and competent? investigating user's impressions of agent's non-verbal behaviours. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*, pages 22–24. ACM.

Biancardi, B., Cafaro, A., and Pelachaud, C. (2017c). Gérer les premières impressions de compétence et de chaleur à travers des indices non verbaux. In *Quatorzièmes Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA 2017)*.

Biancardi, B., Cafaro, A., and Pelachaud, C. (2018). Étude des effets de différents types de comportements non-verbaux sur la perception d'un agent virtuel. *Workshop Affect, Compagnon Artificiel, Interaction (WACAI)*.

Biancardi, B., Mancini, M., Lerner, P., and Pelachaud, C. (under revision). Managing an agent's self-presentational strategies during an interaction. *Frontiers in Robotics and AI - Computational Approaches for Human-Human and Human-Robot Social Interactions*.

Biancardi, B., Wang, C., Mancini, M., Cafaro, A., Chanel, G., and Pelachaud, C. (2019). A computational model for managing impressions of an embodied conversational agent in real-time. In *2019 International Conference on Affective Computing and Intelligent Interaction (ACII)*.

Bickmore, T., Pfeifer, L., and Schulman, D. (2011). Relational agents improve engagement and learning in science museum visitors. In *International Workshop on Intelligent Virtual Agents*, pages 55–67. Springer.

Bickmore, T. W., Vardoulakis, L. M. P., and Schulman, D. (2013). Tinker: a relational agent museum guide. *Autonomous Agents and Multi-Agent Systems*, 27(2):254–276.

Bohra, K. A. and Pandey, J. (1984). Ingratiation toward strangers, friends, and bosses. *The Journal of social psychology*, 122(2):217–222.

Bonaiuto, M., Gnisci, A., and Maricchiolo, F. (2002). Proposta e verifica empirica di una tassonomia dei gesti delle mani nell'interazione di piccolo gruppo. *Giornale italiano di psicologia*, 29(4):777–808.

Bonito, J. A., Burgoon, J. K., and Bengtsson, B. (1999). The role of expectations in human-computer interaction. In *Proceedings of the international ACM SIGGROUP conference on Supporting group work*, pages 229–238. ACM.

Bozeman, D. P. and Kacmar, K. M. (1997). A cybernetic model of impression management processes in organizations. *Organizational behavior and human decision processes*, 69(1):9–30.

Brady, K., Gwon, Y., Khorrami, P., Godoy, E., Campbell, W., Dagli, C., and Huang, T. S. (2016). Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 97–104. ACM.

Brambilla, M., Rusconi, P., Sacchi, S., and Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European journal of social psychology*, 41(2):135–143.

Brambilla, M., Sacchi, S., Rusconi, P., Cherubini, P., and Yzerbyt, V. Y. (2012). You want to give a good impression? be honest! moral traits dominate group impression formation. *British journal of social psychology*, 51(1):149–166.

Brown, B. R. (1970). Face-saving following experimentally induced embarrassment. *Journal of Experimental Social Psychology*, 6(3):255–271.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. (2018). Large-scale study of curiosity-driven learning. In *arXiv:1808.04355*.

Burgoon, J. K. (1993). Interpersonal expectations, expectancy violations, and emotional communication. *Journal of Language and Social Psychology*, 12(1-2):30–48.

Burgoon, J. K., Bonito, J. A., Bengtsson, B., Ramirez Jr, A., Dunbar, N. E., and Miczo, N. (1999). Testing the interactivity model: Communication processes, partner assessments, and the quality of collaborative work. *Journal of management information systems*, 16(3):33–56.

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humpherys, S. L., Moody, G. D., Gaskin, J. E., and Giboney, J. S. (2016). Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies*, 91:24–36.

Burgoon, J. K., Buller, D. B., Hale, J. L., and de Turck, M. A. (1984). Relational messages associated with nonverbal behaviors. *Human Communication Research*, 10(3):351–378.

Burgoon, J. K. and Hale, J. L. (1988). Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communications Monographs*, 55(1):58–79.

Burgoon, J. K. and Walther, J. B. (1990). Nonverbal expectancies and the evaluative consequences of violations. *Human Communication Research*, 17(2):232–265.

Buss, A. H. and Briggs, S. R. (1984). Drama and the self in social interaction. *Journal of Personality and Social Psychology*, 47(6):1310.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*.

Cafaro, A., Vilhjálmsson, H. H., and Bickmore, T. (2016). First impressions in human–agent virtual encounters. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(4):24.

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., Jóhannsdóttir, K. R., and Valgardsson, G. S. (2012). First impressions: Users' judgments of virtual agents' personality

and interpersonal attitude in first encounters. In *International Conference on Intelligent Virtual Agents*. Springer.

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T. W., Heylen, D., and Schulman, D. (2013). First impressions in user-agent encounters: the impact of an agent's nonverbal behavior on users' relational decisions. In *AAMAS*.

Cafaro, A., Wagner, J., Baur, T., Dermouche, S., Torres Torres, M., Pelachaud, C., André, E., and Valstar, M. (2017). The noxi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 350–359. ACM.

Callejas, Z., Ravenet, B., Ochs, M., and Pelachaud, C. (2014). A computational model of social attitudes for a virtual recruiter. *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014*, 1.

Campano, S., Clavel, C., and Pelachaud, C. (2015a). I like this painting too: When an eca shares appreciations to engage users. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems*, pages 1649–1650. International Foundation for Autonomous Agents and Multiagent Systems.

Campano, S., Langlet, C., Glas, N., Clavel, C., and Pelachaud, C. (2015b). An eca expressing appreciations. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 962–967. IEEE.

Camurri, A., Coletta, P., Massari, A., Mazzarino, B., Peri, M., Ricchetti, M., Ricci, A., and Volpe, G. (2004). Toward real-time multimodal processing: Eyesweb 4.0. In *Proceedings of the artificial intelligence and the simulation of behaviour (AISB), 2004 convention: motion. Emotion and cognition*, pages 22–26. Citeseer.

Carrier, A., Louvet, E., and Rohmer, O. (2014). Compétence et agentisme dans le jugement social. *Revue internationale de psychologie sociale*, 27(1):95–125.

Cassell, J. (2000). *Embodied conversational agents*. MIT press.

Cassell, J. et al. (2001). More than just a pretty face: conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, 14(1):55–64.

Castellano, G., Pereira, A., Leite, I., Paiva, A., and McOwan, P. W. (2009). Detecting user engagement with a robot companion using task and social interaction-based features. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 119–126. ACM.

Chen, S., Jin, Q., Zhao, J., and Wang, S. (2017). Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 19–26. ACM.

Choi, A., de Melo, C., Woo, W., and Gratch, J. (2012). Affective engagement to emotional facial expressions of embodied social agents in a decision-making game. *Journal of Visualization and Computer Animation*, 23:331–342.

Chollet, M., Ochs, M., Clavel, C., and Pelachaud, C. (2013). A multimodal corpus approach to the design of virtual recruiters. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*.

Chollet, M., Ochs, M., and Pelachaud, C. (2014). From non-verbal signals sequence mining to bayesian networks for interpersonal attitudes expression. In *International Conference on Intelligent Virtual Agents*, pages 120–133. Springer.

Clavel, C., Cafaro, A., Campano, S., and Pelachaud, C. (2016). Fostering user engagement in face-to-face human-agent interactions: a survey. In *Toward Robotic Socially Believable Behaving Systems-Volume II*, pages 93–120. Springer.

Clavel, C. and Carrión, Z. C. (2016). Sentiment analysis: From opinion mining to human-agent interaction. *IEEE Transactions on Affective Computing*, 7:74–93.

Corrigan, L. J., Peters, C., Küster, D., and Castellano, G. (2016). Engagement perception and generation for social robots and virtual agents. In *Toward Robotic Socially Believable Behaving Systems-Volume I*, pages 29–51. Springer.

Cowie, R. and McKeown, G. (2010). Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme. *SEMAINE Report D6b*.

Cuddy, A. J., Fiske, S. T., and Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149.

Cuddy, A. J., Glick, P., and Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior*, 31:73–98.

De Bruin, E. N. and Van Lange, P. A. (1999). Impression formation and cooperative behavior. *European Journal of Social Psychology*, 29(2-3):305–328.

de Lemus, S., Spears, R., Bukowski, M., Moya, M., and Lupiáñez, J. (2013). Reversing implicit gender stereotype activation as a function of exposure to traditional gender roles. *Social Psychology*.

de Melo, C. M., Carnevale, P. J., Read, S. J., and Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of personality and social psychology*, 106(1):73.

DePaulo, B. M. (1992). Nonverbal behavior and self-presentation. *Psychological bulletin*, 111(2):203.

Dermouche, S. and Pelachaud, C. (2018). From analysis to modeling of engagement as sequences of multimodal behaviors. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., et al. (2014). Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems.

D'Mello, S. K., Olney, A., Williams, C., and Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *Int. J. Hum.-Comput. Stud.*, 70:377–398.

Doherty, K. and Doherty, G. (2018). Engagement in hci: Conception, theory and measurement. *ACM Comput. Surv.*, 51(5):99:1–99:39.

Duchenne, d. B. (1990). The mechanism of human facial expression or an electrophysiological analysis of the expression of the emotions (a. cuthbertson, trans.). *New York: Cam-bridge University Press.(Original work pub-lished 1862)*.

Duval, S. and Wicklund, R. A. (1972). A theory of objective self awareness.

E. Dent, H. (1974). Cultural bias in psychological testing.

Ekman, P. (2002). Facial action coding system (facs). *A human face*.

Ekman, P. and Friesen, W. V. (1974). Nonverbal behavior and psychopathology. *The psychology of depression: Contemporary theory and research*, pages 3–31.

Elliott, C. and Brzezinski, J. (1998). Autonomous agents as synthetic characters. *AI magazine*, 19(2):13–13.

Fiske, S. T., Cuddy, A. J., and Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2):77–83.

Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82(6):878.

Fiske, S. T. and Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in experimental social psychology*, 23:1–74.

Frey, D. (1978). Reactions to success and failure in public and in private conditions. *Journal of Experimental Social Psychology*, 14(2):172 – 179.

Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., and Hagita, N. (2002). Messages embedded in gaze of interface agents - impression management with agent's gaze. In *CHI*.

Glas, N. and Pelachaud, C. (2015a). Definitions of engagement in human-agent interaction. In *International Workshop on Engagment in Human Computer Interaction (ENHANCE)*, pages 944–949.

Glas, N. and Pelachaud, C. (2015b). Politeness versus perceived engagement: an experimental study. *Natural Language Processing and Cognitive Science: Proceedings*, 2014:135.

Glas, N. and Pelachaud, C. (2015c). Topic transition strategies for an information-giving agent. In *European Workshop on Natural Language Generation (ENLG)*, pages 146–155.

Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., and and, A. C. S. (2005). Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Goffman, E. et al. (1978). *The presentation of self in everyday life*. Harmondsworth.

Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., Das, M., and Breazeal, C. (2016). Affective personalization of a social robot tutor for children's second language skills. In *AAAI*, pages 3951–3957.

Gratch, J., DeVault, D., and Lucas, G. (2016). The benefits of virtual humans for teaching negotiation. In *International Conference on Intelligent Virtual Agents*, pages 283–294. Springer.

Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. (2007). Creating rapport with virtual agents. In *International Workshop on Intelligent Virtual Agents*, pages 125–138. Springer.

Gunes, H. and Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1):68–99.

Gunes, H. and Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136.

Heider, F. (2013). *The psychology of interpersonal relations*. Psychology Press.

Hendrick, C. (1968). Averaging vs summation in impression formation. *Perceptual and Motor Skills*, 27(3_suppl):1295–1302.

Heylen, D. et al. (2001). Embodied agents in virtual environments: The aveiro project.

Johnson, W. L. and Rickel, J. (1997). Steve: An animated pedagogical agent for procedural training in virtual environments. *ACM SIGART Bulletin*, 8(1-4):16–21.

Johnson, W. L., Rickel, J. W., Lester, J. C., et al. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial intelligence in education*, 11(1):47–78.

Jones, E. E. and Davis, K. E. (1965). From acts to dispositions the attribution process in person perception. In *Advances in experimental social psychology*, volume 2, pages 219–266. Elsevier.

Jones, E. E. and Harris, V. A. (1967). The attribution of attitudes. *Journal of experimental social psychology*, 3(1):1–24.

Jones, E. E. and Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. *Psychological perspectives on the self*, 1(1):231–262.

Jost, J. and Hunyady, O. (2002). The psychology of system justification and the palliative function of ideology. *European Review of Social Psychology*, 13.

Judd, C. M., James-Hawkins, L., Yzerbyt, V., and Kashima, Y. (2005). Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *Journal of personality and social psychology*, 89(6):899.

Kantharaju, R. B., De Franco, D., Pease, A., and Pelachaud, C. (2018). Is two better than one?: Effects of multiple agents on user persuasion. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 255–262. ACM.

Kapoor, A. and Picard, R. W. (2001). A real-time head nod and shake detector. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–5. ACM.

Katehakis, M. N. and Veinott Jr, A. F. (1987). The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research*, 12(2):262–268.

Kelley, H. H. (1950). The warm-cold variable in first impressions of persons. *Journal of personality*, 18(4):431–439.

Kelley, H. H. (1967). Attribution theory in social psychology. In *Nebraska symposium on motivation*. University of Nebraska Press.

Kervyn, N., Bergsieker, H., and Fiske, S. (2012). The innuendo effect: Hearing the positive but inferring the negative. *Journal of Experimental Social Psychology - J EXP SOC PSYCHOL*, 48.

Kervyn, N., Yzerbyt, V., and Judd, C. M. (2010). Compensation between warmth and competence: Antecedents and consequences of a negative relation between the two fundamental dimensions of social perception. *European Review of Social Psychology*, 21(1):155–187.

Kervyn, N., Yzerbyt, V. Y., Judd, C. M., and Nunes, A. (2009). A question of compensation: the social life of the fundamental dimensions of social perception. *Journal of Personality and Social Psychology*, 96(4):828.

Kim, Y. (2007). Desirable characteristics of learning companions. *International Journal of Artificial Intelligence in Education*, 17(4):371–388.

Kim, Y. and Baylor, A. L. (2006). A social-cognitive framework for pedagogical agents as learning companions. *Educational Technology Research and Development*, 54(6):569–596.

Kopp, S. (2010). Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Communication*, 52:587–597.

Kopp, S., Gesellensetter, L., Krämer, N. C., and Wachsmuth, I. (2005). A conversational agent as museum guide – design and evaluation of a real-world application. In Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., and Rist, T., editors, *Intelligent Virtual Agents*, pages 329–343. Springer Berlin Heidelberg.

Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., Thórisson, K. R., and Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*, pages 205–217. Springer.

Kopp, S. and Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52.

Krenn, B. et al. (2011). Embodied conversational characters: Representation formats for multimodal communicative behaviours. In *Emotion-Oriented Systems*, pages 389–415.

Laban, R. and Lawrence, F. (1979). *Effort: Economy of Human Movement*. Macdonald Evans, London.

Lampel, A. K. and Anderson, N. H. (1968). Combining visual and verbal information in an impression-formation task. *Journal of personality and social psychology*, 9(1):1.

Le Deist, F. D. and Winterton, J. (2005). What is competence? *Human resource development international*, 8(1):27–46.

Leach, C. W., Ellemers, N., and Barreto, M. (2007). Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of personality and social psychology*, 93(2):234.

Leary, M. and Lamphere, R. (1988). Exclusionary self-presentation in a self-presentational dilemma: effects of incongruency between sefl-presentations and target values.

Leary, M. R. and Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological bulletin*, 107(1):34.

Leary, M. R., Robertson, R. B., Barnes, B. D., and Miller, R. S. (1986). Self-presentations of small group leaders: Effects of role requirements and leadership orientation. *Journal of Personality and Social Psychology*, 51(4):742.

Lee, J. and Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In *IVA*.

Lieberman, M. D., Gaunt, R., Gilbert, D. T., and Trope, Y. (2002). Reflexion and reflection: A social cognitive neuroscience approach to attributional inference. In *Advances in experimental social psychology*, volume 34, pages 199–249. Elsevier.

Liew, T. W., Tan, S.-M., and Jayothisa, C. (2013). The effects of peer-like and expert-like pedagogical agents on learners' agent perceptions, task-related attitudes, and learning achievement. *Journal of Educational Technology & Society*, 16(4):275–286.

Lin, X. (2009). Exploration of affect sensing from speech and metaphorical text. In *Edutainment*.

Liu, C., Conn, K., Sarkar, N., and Stone, W. (2008). Online affect detection and robot behavior adaptation for intervention of children with autism. *IEEE transactions on robotics*, 24(4):883–896.

Lombard, M., Ditton, T., Weinstein, L. J., and Temple, J. G. (2009). 1 measuring presence : The temple presence inventory.

Lucas, G. M., Gratch, J., King, A., and Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.

Mancini, M., Biancardi, B., Dermouche, S., Lerner, P., and Pelachaud, C. (2019). An architecture for agent's impression management based on user's engagement. In *International Conference on Intelligent Virtual Agents*. Springer.

Mancini, M., Biancardi, B., Pecune, F., Varni, G., Ding, Y., Pelachaud, C., Volpe, G., and Camurri, A. (2017). Implementing and evaluating a laughing virtual character. *ACM Trans. Internet Techn.*, 17:3:1–3:22.

Mancini, M. and Pelachaud, C. (2008). The fml-apml language. In *Proc. of the Workshop on FML at AAMAS*, volume 8.

Maricchiolo, F., Gnisci, A., Bonaiuto, M., and Ficca, G. (2009). Effects of different types of hand gestures in persuasive speech on receivers' evaluations. *Language and Cognitive Processes*, 24(2):239–266.

Mariooryad, S. and Busso, C. (2015). Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing*, 6(2):97–108.

Marsi, E. and van Rooden, F. (2007). Expressing uncertainty with a talking head in a multimodal question-answering system. In *MOG 2007 Workshop on Multimodal Output Generation*, page 105.

McKeown, G., Curran, W., Wagner, J., Lingenfelser, F., and André, E. (2015). The belfast storytelling database: A spontaneous social interaction database with laughter focused annotation. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 166–172. IEEE.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schröder, M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1).

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

McRorie, M., Sneddon, I., de Sevin, E., Bevacqua, E., and Pelachaud, C. (2009). A model of personality and emotional traits. In *International Workshop on Intelligent Virtual Agents*, pages 27–33. Springer.

Mead, G. H. (1934). *Mind, self and society*, volume 111. Chicago University of Chicago Press.

Mehrabian, A. (1969). Significance of posture and position in the communication of attitude and status relationships. *Psychological Bulletin*, 71(5):359.

Metallinou, A. and Narayanan, S. (2013). Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE.

Miller, D. T. and Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological bulletin*, 82(2):213.

Miller, R. (2010). *Intimate relationships (5e)*.

Moore, C. (2007). Impression formation. *The Blackwell encyclopedia of sociology*.

Morency, L.-P., de Kok, I., and Gratch, J. (2009). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20:70–84.

Nakano, Y. I. and Ishii, R. (2010). Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *IUI*.

Naumann, L. P., Vazire, S., Rentfrow, P. J., and Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and social psychology bulletin*, 35(12):1661–1671.

Nguyen, T.-H. D., Carstensdottir, E., Ngo, N., El-Nasr, M. S., Gray, M., Isaacowitz, D., and Desteno, D. (2015). Modeling warmth and competence in virtual characters. In *International Conference on Intelligent Virtual Agents*, pages 167–180. Springer.

Niewiadomski, R., Demeure, V., and Pelachaud, C. (2010). Warmth, competence, believability and virtual agents. In *International Conference on Intelligent Virtual Agents*, pages 272–285. Springer.

Niewiadomski, R., Hyniewska, S. J., and Pelachaud, C. (2011). Constraint-based model for synthesis of multimodal sequential expressions of emotions. *IEEE Transactions on Affective Computing*, 2:134–146.

Niewiadomski, R., Mancini, M., Baur, T., Varni, G., Griffin, H., and Aung, M. S. (2013). Mmli: Multimodal multiperson corpus of laughter in interaction. In *International Workshop on Human Behavior Understanding*, pages 184–195. Springer.

Nijholt, A. (2002). Embodied agents: A new impetus to humor research.

Nomura, T., Kanda, T., and Suzuki, T. (2006). Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. *Ai & Society*, 20(2):138–150.

Novielli, N. (2010). Hmm modeling of user engagement in advice-giving dialogues. *Journal on Multimodal User Interfaces*, 3(1-2):131–140.

Oden, G. C. and Anderson, N. H. (1971). Differential weighting in integration theory. *Journal of Experimental Psychology*, 89(1):152.

Osherenko, A., André, E., and Vogt, T. (2009). Affect sensing in speech: Studying fusion of linguistic and acoustic features. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6.

Pauchet, A. and Sabouret, N. (2012). Embodied conversational agents and affective computing. 14th European Agent Systems Summer School.

Pecune, F., Cafaro, A., Chollet, M., Philippe, P., and Pelachaud, C. (2014). Suggestions for extending saiba with the vib platform. In *W'shop Architectures and Standards for IVAs, Int'l Conf. Intelligent Virtual Agents*, pages 16–20. Citeseer.

Peeters, G. (1983). Relational and informational patterns in social cognition. *Current issues in European social psychology*, 1:201–237.

Peeters, H. and Lievens, F. (2006). Verbal and nonverbal impression management tactics in behavior description and situational interviews. *International Journal of Selection and Assessment*, 14(3):206–222.

Pelachaud, C. (2009). Modelling multimodal expression of emotion in a virtual agent. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1535):3539–3548.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, 211(2828):42–45.

Peters, C., Asteriadis, S., Karpouzis, K., and de Sevin, E. (2008). Towards a real-time gaze-based shared attention for a virtual agent. In *Workshop on Affective Interaction in Natural Environments (AFFINE), ACM International Conference on Multimodal Interfaces (ICMI'08)*.

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., and Poggi, I. (2005a). A model of attention and interest using gaze behavior. In *International Workshop on Intelligent Virtual Agents*, pages 229–240. Springer.

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., Poggi, I., and Tre, U. R. (2005b). Engagement capabilities for ecas. In *AAMAS'05 workshop Creating Bonds with ECAs*.

Plous, S. (1993). The psychology of judgment and decision making. mcgraw-hill series in social psychology.

Poggi, I. (2007). *Mind, hands, face and body: a goal and belief view of multimodal communication*. Weidler.

Poppe, R., Truong, K. P., and Heylen, D. (2011). Backchannels: Quantity, type and timing matters. In *International Workshop on Intelligent Virtual Agents*, pages 228–239. Springer.

Poppe, R., Truong, K. P., and Heylen, D. (2013). Perceptual evaluation of backchannel strategies for artificial listeners. *Autonomous agents and multi-agent systems*, 27(2):235–253.

Povolny, F., Matejka, P., Hradis, M., Popková, A., Otrusina, L., Smrz, P., Wood, I., Robin, C., and Lamel, L. (2016). Multimodal emotion recognition for avec 2016 challenge. In

Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pages 75–82. ACM.

Ramirez Jr, A. and Wang, Z. (2008). When online meets offline: An expectancy violations theory perspective on modality switching. *Journal of Communication*, 58(1):20–39.

Ravenet, B., Ochs, M., and Pelachaud, C. (2013a). From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In *IVA*.

Ravenet, B., Ochs, M., and Pelachaud, C. (2013b). From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In *International Workshop on Intelligent Virtual Agents*, pages 263–274. Springer.

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., and Trafimow, D. (2002). Inferences about the morality of an aggressor: The role of perceived motive. *Journal of personality and social psychology*, 83(4):789.

Reeves, B. and Nass, C. (1996). *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press.

Reinhard, M.-A., Stahlberg, D., and Messner, M. (2008). Failure as an asset for high-status persons–relative group performance and attributed occupational success. *Journal of Experimental Social Psychology*, 44(3):501–518.

Riggio, R. E. and Friedman, H. S. (1986). Impression formation: The role of expressive behavior. *Journal of Personality and Social Psychology*, 50(2):421.

Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.-P., Ebrahimi, T., Lalanne, D., and Schuller, B. (2015). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30.

Ritschel, H., Baur, T., and André, E. (2017). Adapting a robot's linguistic style based on socially-aware reinforcement learning. In *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on*, pages 378–384. IEEE.

Rosenberg, S., Nelson, C., and Vivekananthan, P. (1968). A multidimensional approach to the structure of personality impressions. *Journal of personality and social psychology*, 9(4):283.

Rosenberg-Kima, R. B., Baylor, A. L., Plant, E. A., and Doerr, C. E. (2008). Interface agents as social models for female students: The effects of agent visual presence and appearance on female students' attitudes and beliefs. *Computers in Human Behavior*, 24(6):2741–2756.

Rosenfeld, H. M. (1966). Approval-seeking and approval-inducing functions of verbal and nonverbal responses in the dyad. *Journal of Personality and Social Psychology*, 4(6):597.

Rummery, G. A. and Niranjan, M. (1994). *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, England.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Ruttkay, Z., Noot, H., and ten Hagen, P. J. W. (2003). Emotion disc and emotion squares: Tools to explore the facial expression space. *Comput. Graph. Forum*, 22:49–54.

Salam, H., Celiktutan, O., Hupont, I., Gunes, H., and Chetouani, M. (2017). Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access*, 5:705–721.

Schlenker, B. R. (1980). *Impression management*. Brooks/Cole Publishing Company Monterey, CA.

Schlenker, B. R. (1985). Identity and self-identification. *The self and social life*, 65:99.

Schlenker, B. R. and Darby, B. W. (1981). The use of apologies in social predicaments. *Social psychology quarterly*, pages 271–278.

Schneider, D. J. (1969). Tactical self-presentation after success and failure. *Journal of Personality and Social Psychology*, 13(3):262.

Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B. W., de Sevin, E., Valstar, M. F., and Wöllmer, M. (2015). Building autonomous sensitive artificial listeners (extended abstract). *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 456–462.

Schuller, B. W., Rigoll, G., and Lang, M. K. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:I–577.

Scotia, N. (2010). Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*, 19:227.

Shelton, C. R. (2001). Balancing multiple sources of reward in reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1082–1088.

Sidner, C. L. and Dzikovska, M. (2005). A first experiment in engagement for human-robot interaction in hosting activities. In *Advances in natural multimodal dialogue systems*, pages 55–76. Springer.

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Artif. Intell.*, 166:140–164.

Singh, R. and Teoh, J. B. P. (2000). Impression formation from intellectual and social traits: Evidence for behavioural adaptation and cognitive processing. *British Journal of Social Psychology*, 39(4):537–554.

Skowronski, J. J. and Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of personality and social psychology*, 52(4):689.

Smith-Lovin, L. and Brody, C. (1989). Interruptions in group discussions: The effects of gender and group composition. *American Sociological Review*.

Snyder, C. R., Higgins, R. L., and Stucky, R. J. (1983). *Excuses: Masquerades in search of grace*. Number 341. John Wiley & Sons.

Straßmann, C., von der Pütten, A. R., Yaghoubzadeh, R., Kaminski, R., and Krämer, N. (2016). The effect of an intelligent virtual agent's nonverbal behavior with regard to dominance and cooperativity. In *International Conference on Intelligent Virtual Agents*, pages 15–28. Springer.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., Lane, C., Morie, J., Aggarwal, P., Liewer, M., Chiang, J.-Y., Gerten, J., Chu, S., and White, K. (2010). Ada and grace: Toward realistic and engaging virtual museum guides. In Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., and Safonova, A., editors, *Intelligent Virtual Agents*, pages 286–300. Springer Berlin Heidelberg.

Ter Maat, M., Truong, K. P., and Heylen, D. (2010). How turn-taking strategies influence users' impressions of an agent. In *IVA*, volume 6356, pages 441–453. Springer.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of applied psychology*, 4(1):25–29.

Traum, D. R., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., Leuski, A., Noren, D., and Swartout, W. R. (2012). Ada and grace: Direct interaction with museum visitors. In *IVA*.

Truong, K. P., Poppe, R., and Heylen, D. (2010). A rule-based backchannel prediction model using pitch and pause information. In *Eleventh Annual Conference of the International Speech Communication Association*.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309.

Urbain, J. et al. (2009). Avlaughtercycle: An audiovisual laughing machine. In *Proceedings of the 5th International Summer Workshop on Multimodal Interfaces*, pages 79–87.

Valstar, M., Baur, T., Cafaro, A., Ghitulescu, A., Potard, B., Wagner, J., André, E., Durieu, L., Aylett, M., Dermouche, S., et al. (2016). Ask alice: an artificial retrieval of information agent. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 419–420. ACM.

van Waterschoot, J., Bruijnes, M., Flokstra, J., Reidsma, D., Davison, D., Theune, M., and Heylen, D. (2018). Flipper 2.0: A pragmatic dialogue engine for embodied conversational agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 43–50. ACM.

Veletsianos, G. (2010). Contextually relevant pedagogical agents: Visual appearance, stereotypes, and first impressions and their impact on learning. *Computers & Education*, 55(2):576–585.

Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N. E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., et al. (2007). The behavior markup language: Recent developments and challenges. In *International Workshop on Intelligent Virtual Agents*, pages 99–111. Springer.

Wang, C., Chanel, G., Biancardi, B., Mancini, M., and Pelachaud, C. (submitted). Automatic impression detection and a use case with an embodied conversational agent. In *21th ACM International Conference on Multimodal Interaction*. ACM.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., and Collins, H. (2005). The politeness effect: Pedagogical agents and learning gains. In *AIED*.

Watkins, C. J. C. H. (1989). Learning from delayed rewards.

Willis, J. and Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7):592–598.

With, S. and Kaiser, S. (2017). Sequential patterning of facial actions in the production and perception of emotional expressions with ,.

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology*, 67(2):222.

Wojciszke, B. (2005). Morality and competence in person-and self-perception. *European review of social psychology*, 16(1):155–188.

Wojciszke, B. and Abele, A. E. (2008). The primacy of communion over agency and its reversals in evaluations. *European Journal of Social Psychology*, 38(7):1139–1147.

Wojciszke, B., Bazinska, R., and Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, 24(12):1251–1263.

Yang, Y.-H. and Chen, H. H. (2011). Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):762–774.

Ybarra, O., Chan, E., and Park, D. (2001). Young and old adults' concerns about morality and competence. *Motivation and Emotion*, 25(2):85–100.

Ybarra, O., Chan, E., Park, H., Burnstein, E., Monin, B., and Stanik, C. (2008). Life's recurring challenges and the fundamental dimensions: An integration and its implications for cultural differences and similarities. *European Journal of Social Psychology*, 38(7):1083–1092.

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting, 1970*, pages 567–578.

Yzerbyt, V., Provost, V., and Corneille, O. (2005). Not competent but warm... really? compensatory stereotypes in the french-speaking world. *Group Processes & Intergroup Relations*, 8(3):291–308.

Yzerbyt, V. Y., Kervyn, N., and Judd, C. M. (2008). Compensation versus halo: The unique relations between the fundamental dimensions of social judgment. *Personality and Social Psychology Bulletin*, 34(8):1110–1123.

Zhao, R., Sinha, T., Black, A., and Cassell, J. (2016). Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 381–392.