

**Dimensions affectives et sociales
dans les interactions parlées avec des (ro)bots :
enjeux technologiques, pluridisciplinaires
et éthiques**

par Laurence Devillers

Introduction

Le dialogue en interaction sociale implique naturellement l'échange de contenu linguistique, mais aussi et peut-être de façon plus importante l'expression de relations interpersonnelles, aussi bien que d'expressions d'émotion, de compréhension, d'intérêt, etc. En tant qu'outil de communication, le langage a connu des bouleversements inédits liés notamment à de nouveaux modes de communication très vite adoptés / imposés, via des interfaces numériques, en situation de travail, d'apprentissage, etc. L'utilisation des modalités linguistiques mais aussi non linguistiques (paralinguistiques et extralinguistiques) va faciliter la réalisation de tâches partagées entre agent et humain en exploitant les apports relatifs au contexte linguistique et non linguistique pour échanger des informations dans une communication située.

Dans un avenir proche, la robotique sociale et les agents conversationnels d'assistance (*chatbots*) seront utilisés pour certains aspects de l'éducation, du travail, du soin, en automatisant la surveillance, l'entraînement, la motivation et leur aspect « compagnon » pour les personnes âgées, les enfants, les personnes

handicapées ou même les individus souffrant de phobies sociales ou de maladies dégénératives. Les (ro)bots « affectifs », comme les enceintes vocales qui sont déjà au sein de nombreux foyers, seront de plus en plus utilisés dans la société, pourraient devenir un moyen d'influence des individus. L'utilisation de *chatbots* et robots personnalisés grâce à la détection des émotions pourrait amplifier les phénomènes de manipulation et d'influence notamment pour des personnes vulnérables comme les enfants et les personnes âgées. L'influence des affects sur la prise de décision est importante. Les nudges, sortes de manipulations douces, issues des sciences comportementales consistent à inciter les individus à modifier leur comportement sans les contraindre, en utilisant leurs biais cognitifs et leurs réactions émotionnelles. Les travaux sur le nudging prennent leurs sources dans les travaux de R. Thaler, prix Nobel d'économie en 2017 (Thaler et Sunstein 2008).

Le projet pluridisciplinaire Bad-Nudge Bad-Robot (Devillers 2018) dans l'institut Dataia (Paris-Saclay), qui réunit informaticiens, linguistes, économistes et juristes, veut répondre à plusieurs verrous de recherche sur le nudge dans l'interaction humain-machine. Il s'inscrit pleinement dans la Chaire HUMAINE : HUman-MACHine Affective INTERaction & Ethics (Devillers 2019) au Limsi-CNRS qui porte sur l'étude des comportements des machines et de la co-existence Humain-Machine. Cette chaire combine l'aspect de recherche scientifique en intelligence artificielle avec la mise en œuvre d'une méthodologie innovante pour évaluer et améliorer l'éthique de l'interaction affective homme-machine, malgré l'opacité des systèmes concernés. Elle met en avant une forte collaboration inter et pluridisciplinaire entre l'informatique affective, l'économie comportementale, la linguistique et le traitement du langage naturel. Des premières expérimentations fondées sur la théorie des jeux ont été menées dans le projet Bad-Nudge Bad-Robot pour montrer l'influence des nudges en situation réelle avec des enfants.

Les enfants apprennent en expérimentant le monde. Ces agents et robots sont dits « intelligents » car ils peuvent également apprendre. Pour un robot ou un agent, la tâche est extrêmement difficile car il n'a ni instinct, ni intentions pour prendre des décisions. Par contre, il peut imiter l'humain. Ces machines ne ressen-

tent rien, n'ont pas de conscience malgré le fait qu'elles pourront dire de façon empathique « je suis triste pour toi ! » Les machines seront de plus en plus autonomes, « bavardes » et douées d'émotions grâce à des programmes d'intelligence artificielle sophistiqués mais elles ne seront pas pour autant capables de sentiments, de créativité et d'imagination tels qu'on l'entend pour les humains. La machine ne peut pas se sentir « triste » car elle est dénuée d'un corps qui ressent et de conscience phénoménale. Elle ne peut pas d'ailleurs « comprendre » au sens humain le concept de tristesse. Malgré cela, tous ces outils artificiellement intelligents deviennent de formidables alliés dans une interaction complémentaire avec les humains. La robotique sociale et affective veut créer des robots compagnons, censés apporter une aide thérapeutique, d'assistance ou encore de surveillance. Il est nécessaire qu'une réflexion éthique pluridisciplinaire soit combinée avec le développement technologique des robots et *chatbots*, ainsi qu'une surveillance du comportement des machines (Rahwan *et alii* 2019), pour assurer l'harmonie et l'acceptabilité de leur relation avec les êtres humains.

1. Agent conversationnel

Un agent conversationnel, appelé communément *chatbot*, est un programme informatique qui interagit avec son utilisateur dans la langue naturelle de celui-ci. Ces termes regroupent tant les agents vocaux que les *chatbots* écrits. Le plus souvent, un agent conversationnel ne constitue pas une entité indépendante, mais il est intégré au sein d'un système ou d'une plate-forme numérique, comme un smartphone ou une enceinte vocale. Sur le plan de l'apparence visuelle, les *chatbots* peuvent aussi être intégrés à un agent conversationnel animé, représenté en deux ou trois dimensions sur un écran, voire faire partie d'un robot social, y compris humanoïde. Le dialogue avec l'utilisateur ne représente alors qu'une des fonctions du système global. L'histoire des agents conversationnels prend ses origines dans le jeu de l'imitation d'Alan Turing. La compréhension du langage intéresse Turing dans la mesure où elle se manifeste à travers des réponses qui paraissent intelligibles et sensées à un examinateur (« test de Turing »). Dès 1991, un concours annuel est organisé afin de sou-

tenir le développement de *chatbots* capables de passer le test de Turing. Le premier agent conversationnel de l'histoire de l'informatique est le programme *Eliza* de Joseph Weizenbaum, qui est aussi l'un des premiers leurres conversationnels. *Eliza* simule un dialogue écrit avec un psychothérapeute rogérien en reformulant tout simplement la plupart des répliques de l'utilisateur « patient » sous forme de questions. Aujourd'hui, l'expression « effet *Eliza* » désigne la tendance à assimiler de manière inconsciente le dialogue avec un ordinateur à celui avec un être humain.

La conception et le fonctionnement d'un agent conversationnel se divisent en plusieurs modules de traitement automatique du langage naturel (TALN) : schématiquement, un *chatbot* peut inclure des modules de reconnaissance de la parole (pour les agents conversationnels vocaux), de traitement sémantique (hors et en contexte), de gestion de l'historique du dialogue, de gestion des stratégies de dialogue, d'accès aux ontologies, de gestion des accès aux connaissances externes (base de données ou internet), de génération de langage et de synthèse de la parole (pour les agents conversationnels vocaux). Un agent conversationnel suit des règles décidées et transposées en code par des concepteurs humains ou obtenues par apprentissage. Les *chatbots* apprenants, par exemple *XiaoIce* de Microsoft Chine (Zhou 2020), sont aujourd'hui encore assez rares parmi les applications commercialisées.

2. L'informatique affective

Pendant longtemps, philosophes et scientifiques ont opposé raison et émotions. Aujourd'hui, l'évolution des connaissances scientifiques grâce aux neurosciences, montre que les formes émotionnelles sont nécessaires au fonctionnement cognitif, à la mémorisation et à l'apprentissage. Le neuroscientifique Antonio Damasio (Damasio 2003) apporte une vision nouvelle sur la manière dont les émotions se manifestent dans les interrelations étroites qu'entretiennent le corps et le cerveau dans la perception des objets. La notion de « corps » est centrale : l'homme ne se réduit pas à une pensée, à une conscience de soi ; il est aussi un corps, par l'intermédiaire duquel il se trouve dans le monde. L'empathie est par exemple un des piliers des relations humaines et des rapports humains. C'est une réponse émotionnelle à une

situation bien particulière. Pouvoir ressentir une émotion appropriée en réponse à celle exprimée par l'autre en distinguant bien, l'émotion de l'autre et sa propre émotion, est une forme de pensée qui utilise la mémoire et la sensorialité. L'homme possède une capacité particulière à se projeter. Chez l'enfant, l'empathie affective apparaît à un an, l'empathie cognitive apparaît plus tard vers quatre ans et demi. Une empathie plus « mature » incluant le sens de la réciprocité et permettant la construction du sens moral et de la justice est présente entre huit et douze ans.

L'informatique affective est née en 1997 des travaux au MIT de Rosalind Picard (Picard 1997) et regroupe trois technologies : la reconnaissance des émotions des humains, le raisonnement et la prise de décision en utilisant ces informations, la génération d'expressions émotionnelles. Ce domaine est par essence pluridisciplinaire. La reconnaissance des messages sociaux véhiculés par les visages et les voix, et en particulier les expressions émotionnelles, est un élément indispensable à la communication avec les humains et à l'insertion dans toutes les sociétés. De façon consensuelle, l'émotion est définie comme une réaction à un événement / situation réelle ou imaginaire comprenant plusieurs facettes ou composantes. Trois composantes sont généralement acceptées comme constituants essentiels de la réaction émotionnelle. Il s'agit du sentiment subjectif (vécu émotionnel), de la réaction physiologique (tremblement, rythme cardiaque accéléré) et de l'expression émotionnelle (faciale, vocale, gesturale ou posturale). Dans un contexte de communication, les expressions sont transformées en fonction d'un ensemble de règles socio-culturelles. Ces règles varient d'une culture ou d'un groupe social à l'autre, dans des contextes « objectivement » similaires. Des situations sociales peuvent exiger la suppression de certaines expressions alors que d'autres situations au contraire exigent de montrer voire d'exagérer des expressions spécifiques. Masquer une expression spontanée qui ne serait pas désirable dans un contexte social donné, est également possible. Le contexte social peut influencer les expressions émotionnelles en fonction de la position et des objectifs de l'émetteur dans la situation. Dans une situation de communication un individu peut utiliser ses expressions émotionnelles de manière à influencer – plus ou moins consciemment et plus ou moins volontairement – les réactions de ses interlocuteurs.

Ces technologies de simulations affectives commencent à arriver dans les objets autour de nous. Une fois entrée dans l'usage quotidien, l'informatique affective a le potentiel de changer massivement notre façon d'interagir avec les dispositifs informatiques et robotiques : ils pourront répondre de manière plus appropriée à nos émotions et à nos humeurs, et montrer des signes d'empathie par imitation. L'informatique affective, si elle devient robuste, a également le potentiel de changer massivement les soins de santé mentale, une fois que les systèmes informatiques seront capables de surveiller notre bien-être ou une dépression potentielle, ou, juste à titre d'exemple, le développement de nos enfants. Toutefois, dans la plupart de ces applications, sinon toutes, une évaluation fiable de l'affect et du comportement affectif est essentielle ainsi qu'une évaluation en continu de ces systèmes. Une percée majeure dans ce domaine – comme cela a été le cas pour de nombreux problèmes connexes de traitement intelligent des signaux – est venue avec l'avènement et l'utilisation croissante de l'apprentissage profond. Si cette approche a permis un nombre impressionnant de succès dans l'amélioration des performances, il est notable que les approches n'intègrent pas encore assez d'informations contextuelles pour pouvoir reconnaître et interpréter de façon fiable les émotions, tout particulièrement quand la tâche est non spécifiée.

3. L'interaction affective avec les machines

Établir une interaction affective avec les robots n'est plus un rêve d'auteur de science-fiction mais une thématique émergente pour de nombreuses équipes de recherche. Au Limsi-CNRS, nous travaillons à donner aux robots la capacité de reconnaître les émotions et d'être empathiques, pour qu'ils aident au mieux leur utilisateur. Nous leur apprenons à dialoguer et à analyser les émotions à partir d'indices verbaux et non verbaux (rires, par exemple) afin d'adapter leurs réponses (Devillers *et alii* 2014).

Les robots sociaux vont partager notre espace, habiter nos maisons, nous aider dans notre travail et notre vie quotidienne et, également, partager avec nous une certaine histoire. Pourquoi ne pas les doter d'humour machine ? L'humour joue un rôle capital dans les relations sociales : il amortit le stress, met en confiance,

crée une connivence entre les interlocuteurs. Si vous êtes seul et malheureux, le robot pourrait plaisanter pour vous reconforter ; si vous êtes en colère, il pourrait vous aider à relativiser, dire que la situation n'est pas si terrible. Il pourrait également faire preuve d'autodérision s'il commet des erreurs et s'en rend compte !

Comment ces robots « empathiques » sont-ils accueillis ? Pour le savoir, il est important de mener des études perceptives sur l'interaction humain-machine. Le Limsi-CNRS a conduit de nombreux tests en laboratoire et en Ehpad avec des personnes âgées, ou dans des centres de rééducation avec l'association Approche, dans le cadre du projet BPI Romeo2, piloté par *Softbank robotics* (Pandey *et alii* 2014). Créée en 1991, l'association Approche a pour principale mission de promouvoir les nouvelles technologies (robotique, électronique, domotique, technologies de l'information et de la communication...) au service des personnes en situation de handicap quels qu'en soient l'âge et le milieu de vie. Nous avons également mené des études autour de scénarios de la vie de tous les jours et de jeux avec l'équipe du Pr Anne-Sophie Rigaud au Living Lab de l'hôpital Broca (Garcia *et alii* 2017). Toutes ces expériences ont montré que les robots étaient plutôt bien acceptés par les patients lorsqu'ils ont le temps de les expérimenter. De nombreux travaux de recherche sont encore nécessaires pour étudier la relation à long terme avec ces machines.

Les nudges, suggestions indirectes qui peuvent affecter le comportement et la prise de décision, sont considérés dans le contexte des machines conversationnelles. (Thaler et Sunstein 2008, 2012) ont souligné le fait que le *nudging* était une tactique visant à modifier subtilement le comportement d'une personne, sans restreindre son choix. En effet, le nudging opère principalement par le biais du système affectif ou en exploitant des biais cognitifs communs (par exemple, l'attention, la mémoire, la paresse). Si les nudges peuvent être déployés pour encourager les individus à exprimer des comportements qui ont des avantages pour la communauté, un nudge pourrait avoir des conséquences imprévues pour les personnes dont les antécédents n'ont pas été bien pris en compte dans le développement du système. De même, les nudges peuvent encourager des comportements ayant des effets imprévus à long terme, qu'ils soient positifs ou négatifs, pour l'individu et/ou la société. L'effet de l'IA sur une personne,

comme l'érosion ou l'encouragement potentiel de la liberté individuelle, ou l'expression de comportements qui profitent à autrui, doit être bien caractérisé dans la conception de l'IA. Tous nos choix sont influencés par l'environnement de choix. Les concepteurs de cet environnement de choix peuvent donc utiliser l'heuristique et les biais humains (confirmation, effet de leurre, ancrage, etc.) pour manipuler l'environnement afin de guider subtilement le comportement des utilisateurs en les poussant doucement vers certains choix (Schneider *et alii* 2018). Le nudging numérique peut être considéré comme un sous-ensemble de l'informatique de persuasion.

Nos travaux portent sur la détection des nudges dans les dialogues parlés à partir de l'interaction homme-machine et de la conversation homme-homme. Le mimétisme audio (comme le mimétisme facial) est une marque d'engagement entre deux personnes dans une interaction sociale. Le décodage conjoint des canaux lexicaux et audio pourra aider à désambiguïser le contenu sémantique mais aussi à calculer une métrique d'alignement entre deux personnes en utilisant le signal dans les disfluences et autres caractéristiques audio qui caractérisent la forme la plus humaine de transmission et de génération d'informations : la communication conversationnelle. En particulier, l'utilisation du canal audio dans la segmentation de la conversation parlée pourrait produire des unités plus flexibles, plus robustes et plus significatives. Parmi les nombreux défis à relever, nos travaux soulèvent la question de l'utilisation éthique pour des populations sensibles telles que les enfants ou les personnes âgées qui ne sont pas suffisamment conscients de la technologie et des intentions qui se cachent derrière ces dispositifs.

4. Éthique des systèmes robotiques

Les robots sociaux et affectifs soulèvent de nombreuses questions éthiques, juridiques et sociales. Qui est responsable en cas d'accident : le fabricant, l'acheteur, le thérapeute, l'utilisateur ? Comment réguler leur fonctionnement ? Faut-il intégrer des règles morales dans leur programmation ? Contrôler leur utilisation par des permis ? Pour quelles tâches souhaitons-nous créer ces entités artificielles ? Comment préserver notre intimité, nos données

personnelles ? Tout système doit être évalué avant d'être mis dans les mains de son utilisateur (Dubuisson et Devillers 2015). Comment évaluer une intelligence artificielle qui apprend des humains et s'adapte à eux, ou qui apprend seule ? Peut-on prouver qu'elle se cantonnera aux fonctions pour lesquelles elle a été conçue, qu'elle ne pas dépassera pas les limites fixées ? Les données que la machine exploite pour son apprentissage la dirigent vers certaines actions. Qui supervisera la sélection de ces données ?

Ces questions prégnantes ne sont évoquées que depuis peu (Devillers 2017). Les progrès spectaculaires du numérique permettront un jour d'améliorer le bien-être des personnes, à condition de réfléchir non à ce que nous pouvons en faire, mais à ce que nous voulons en faire. C'est pourquoi la plus importante association professionnelle internationale du numérique, l'institut des ingénieurs électriciens et électroniciens (IEEE), organisation savante mondiale, a lancé une initiative pour réfléchir à une éthique relative aux systèmes autonomes. Dans la foulée, une douzaine de groupes de travail sur les normes et les standards ont émergé (Fagella 2019), notamment sur le nudging des robots (manipulation incitative) (P7008). L'utilisation d'applications d'IA permettant de donner des coups de coude (*nudges*) à des ordinateurs, des *chatbots* ou des robots est en constante augmentation dans le monde, tant dans le secteur privé que public, par exemple pour la santé, le bien-être, les applications éducatives, la politique publique gouvernementale pour les impôts, le marketing pour le secteur privé, etc. L'IA offre une formidable opportunité non seulement d'augmenter l'efficacité et de réduire les coûts, mais aussi de construire un processus de nudging avec des règles responsables et éthiques. Déployer l'IA sans l'ancrer à une conformité solide et à des valeurs fondamentales peut exposer à des risques importants, notamment des problèmes de confidentialité des données, de santé et de sécurité. L'impact des conséquences involontaires de l'évolution de la technologie peut entraîner des dommages pour la réputation et, dans le pire des cas, la perte de la confiance des clients et de la valeur des actionnaires. Lors du développement de systèmes robotiques, intelligents et autonomes une norme peut être utilisée comme ligne directrice, et comme référence des « types de stratégies de nudge positif » pour per-

mettre une communication claire et précise entre les membres des différentes communautés qui incluent la robotique, les systèmes intelligents et autonomes, l'éthique et les domaines d'expertise corrélés. Il existe déjà un robot domestique pour inciter les personnes âgées à rester socialement connectées et actives : ElliQ, un robot compagnon pour les personnes âgées, vise à promouvoir l'activité et à lutter contre la solitude en les poussant à participer à des activités numériques et physiques. Des applications comme celle-ci pourraient bénéficier de l'utilisation de normes autour de leurs intentions d'influencer le comportement humain. À l'avenir, la transparence de ces changements de comportement (et les moyens de les encourager) pourrait être remise en question, et les entreprises ayant des normes claires pourraient être plus enclines à gagner la confiance de leurs clients et à éviter les problèmes juridiques. Il faut éviter un déficit de confiance mais également une confiance trop aveugle dans les programmes d'intelligence artificielle. Un certain nombre de valeurs éthiques sont importantes : la déontologie et responsabilité des concepteurs, l'émancipation des utilisateurs, l'évaluation, la transparence, l'explicabilité, la loyauté, et l'équité des systèmes, enfin l'étude de la co-évolution humain-machine (la machine s'adaptera à l'humain et l'humain à la machine).

Conclusion

Les nouveaux usages des robots sociaux, des agents conversationnels et, plus généralement, des environnements numériques dits « intelligents » dans des domaines aussi divers que la santé, l'éducation, l'assurance, les transports ou l'économie reflètent une phase de changement significatif dans les relations homme-machine, qui devrait faire l'objet d'une grande attention. Comment l'Homme va-t-il co-apprendre, co-crée et co-adapter avec la Machine ? Notamment, comment les personnes vulnérables seront-elles protégées contre les menaces potentielles de la machine ? Nous avons besoin de démystifier, de former à l'intelligence artificielle et de remettre au centre de la conception de ces systèmes robotiques, les valeurs de l'humain. L'intelligence artificielle peut apporter de meilleurs diagnostics en santé, des outils de stimulation, de détection des comportements anormaux, une

meilleure assistance notamment pour le handicap, ou la perte d'autonomie. Il est nécessaire de développer des cadres éthiques pour les robots sociaux et affectifs notamment en santé et de comprendre le niveau de complémentarité humain-machine (Devillers 2020).

La chaire HUMAINE vise à étudier ces interactions et relations afin d'auditer et de mesurer l'influence potentielle des systèmes intelligents et affectifs sur l'homme, et enfin d'aller vers une conception des « systèmes éthiques » *by design* et de proposer des mesures d'évaluation. Dans ce but, le travail scientifique prévu se concentre sur la détection des émotions sociales dans la voix humaine, et sur l'étude des nudges audio et parlés, destinés à induire des changements dans le comportement de l'interlocuteur humain. Ces travaux devraient être complétés par des études expérimentales (à long terme, influence de l'homme sur la machine, etc.) pour évaluer les aspects éthiques et la confiance dans les machines, ainsi que par la démystification de ces technologies auprès du grand public qui tend naturellement vers l'anthropomorphisme. L'importance de ce projet réside également dans la variété de ses applications sociétales (des soins aux personnes âgées et vulnérables, à l'économie, et à l'éducation) et de la recherche menée en pluridisciplinarité.

Bibliographie

- DAMASIO Antonio, 2003, *Spinoza avait raison. Le cerveau de la tristesse, de la joie et des émotions*, Paris, Odile Jacob.
- DEVILLERS Laurence, 2017, *Des robots et des hommes : mythes, fantasmes et réalité*, Paris, Plon.
- DEVILLERS Laurence, 2018, *Bad nudge Bad Robot Project*, en ligne.
- DEVILLERS Laurence, 2019, chaire Humaine, CNRS-Limsi, en ligne.
- DEVILLERS Laurence, 2020, *Les robots émotionnels : santé, surveillance, sexualité... et l'éthique dans tout cela ?*, Paris, L'Observatoire.
- DEVILLERS Laurence, TAHON Marie, SEHILI Mohamed et DELABORDE Agnès, 2014, « Détection des états affectifs lors d'interactions parlées : robustesse des indices non verbaux », *TAL*, vol. 55 fasc. 2.
- DEVILLERS Laurence, TAHON Marie, SEHILI Mohamed and DELABORDE Agnès, 2015, "Inference of human beings' emotional states from speech in human-robot interactions", *International Journal of Social Robotics*, vol. 7 n° 4, p. 451-463.

- DUBUISSON-DUPLESSIS Guillaume et DEVILLERS Laurence, 2015, "Towards the consideration of dialogue activities in engagement measures for human-robot social interaction", *International Conference on Intelligent Robots and Systems, Designing and Evaluating Social Robots for Public Settings Workshop*, p. 19-24.
- FAGELLA Daniel, 2019, « Initiative IEEE standards et normes », en ligne.
- GARCIA Mélanie, BECHADE Lucide, Guillaume DUPLESSIS-DUBUISSON, Gabrièle PITTARO et Laurence DEVILLERS, 2017, "Towards metrics of Evaluation of Pepper robot as a Social Companion for Elderly People", *International Workshop on Spoken Dialog System Technology*.
- KUMAR PANDEY Amit, GELIN Rodolphe *et al.*, 2014, "Ethical considerations and feedback from social human-robot interaction with elderly people", *AIC*.
- LI ZHOU Di Li and SHUM Heung-Yeung, 2020, "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot", *Computational Linguistics*, 46 (1), p. 53-93.
- PICARD Rosalind W., 1997, *Affective Computing*, Cambridge (MA), MIT Press.
- RAHWAN Iyad, Manuel CEBRIAN *et al.*, 2019, "Machine behavior", *Nature*, 568, p. 477-486.
- SCHNEIDER Christoph, WEINMANN Marcus and VON BROCKE Jam, 2016, "Digital Nudging, *Business & Information Systems Engineering*, 58 (6), p. 433-436.
- THALER Richard and SUNSTEIN Cass R., 2008, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, New Haven (CT), Yale University Press.
- THALER Richard et SUNSTEIN Cass R., 2012, *Nudge : émotions, habitudes, comportements: comment inspirer les bonnes décisions ?*, Paris, Vuibert.