Des interfaces traditionnelles hommes-machines aux machines empathiques : vers une coadaptation humain-machine

Par **Laurence DEVILLERS**
Professeur en IA à Sorbonne Université/LIMSI-CNRS, membre du CNPEN et du GPAI sur le futur du travail

Introduction

Pour comprendre les particularités des robots empathiques, rappelons qu'un robot est caractérisé par trois composantes en interaction : il recueille des données grâces à ses capteurs, il les interprète grâce à ses programmes, et il peut bouger et agir sur son environnement. De plus, le robot peut avoir une apparence anthropomorphe et une capacité d'interaction langagière. On désigne par *chatbot* ou « agent conversationnel » un système de dialogue homme-machine. La robotique sociale tend à créer des robots dotés de capacités sociales, entre autres la capacité de dialoguer, qui pourraient prendre des rôles de substituts pour certaines tâches dans la société (Dumouchel et Damasio, 2016). Enfin, ces robots sociaux peuvent être dotés d'empathie artificielle.

L'empathie est une réponse émotionnelle à une situation bien particulière. C'est un trait de personnalité qui est de pouvoir ressentir une émotion appropriée en réponse à celle exprimée par une autre personne, et de bien distinguer l'émotion de l'autre (votre enfant a mal) de votre émotion (vous avez mal pour lui, mais vous ne sentez pas de souffrance physique). 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.

Les robots et les *chatbots* comme Google Home ou Siri sur notre téléphone pourraient nous donner l'illusion que les machines sont empathiques, si on leur donne la capacité d'identifier les émotions de leurs interlocuteurs humains, de raisonner à partir des émotions détectées, et de générer par l'expressivité faciale, les gestes, les postures et l'expressivité acoustique des émotions. La détection des émotions d'une personne peut amener la machine à changer de stratégie de réponse. Elle peut ainsi répondre « je suis triste aussi » à quelqu'un, lorsqu'elle a détecté de la tristesse. Cette tromperie peut nous amener à croire à l'empathie des robots. Il ne s'agit pas véritablement d'empathie, car ces machines n'ont pas de « conscience phénoménale », c'est-à-dire d'expériences caractérisant le « vécu » ou le « ressenti » d'une personne.

La conscience phénoménale, contrairement à la conscience définie comme « cognition », est associée à une expérience qualitative, telle que le sentiment de plaisir ou de douleur, de chaud ou de froid, etc. La conscience phénoménale n'est donc pas réductible aux conditions physiques ou physiologiques de son apparition, et elle est indissociable de la subjectivité de l'individu. Il y a ainsi une grande différence entre les descriptions scientifiques de la conscience, qui font référence au comportement ou au fonctionnement du cerveau publiquement observables, et la conscience phénoménale propre au sujet.

Le neuroscientifique Antonio Damasio a apporté 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 (Damasio, 1994). 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. Il faut lever les voiles du mystère des sciences comme l'intelligence artificielle, les neurosciences cognitives et l'intelligence affective pour mieux utiliser dans la société ces artefacts, robots ou agents conversationnels. La modélisation informatique des affects amène à se poser la question des conséquences sociétales de vivre dans un quotidien environné d'objets pseudo-affectifs (Devillers, 2017). Des principes éthiques imposant par exemple de distinguer les agents artificiels pourraient être envisagés.

L'*"affective computing"*

Le domaine de l'"*affective computing*" prend ses sources dans les travaux de Rosalind Picard (Picard, 1997), au MIT en 1997, et regroupe trois technologies : la reconnaissance des émotions des humains, le raisonnement et la prise de décision en utilisant les informations recueillies, et 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, une situation, réel(le) ou imaginaire, comprenant plusieurs facettes ou composantes. Trois composantes sont généralement acceptées comme constituantes essentielles de la réaction émotionnelle. Il s'agit du sentiment subjectif (vécu émotionnel), de la réaction physiologique et de l'expression émotionnelle (faciale, vocale 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 (in)consciemment et plus ou moins (in)volontairement – les réactions de ses interlocuteurs.

Les machines vont de plus en plus interagir vocalement avec nous dans la vie de tous les jours. Les agents conversationnels et les robots sociaux peuvent déjà embarquer des systèmes de détection, de raisonnement et de génération d'expressions affectives qui, même avec des erreurs importantes, peuvent interagir avec nous. Ils envahissent maintenant nos sphères privées. On dénombre aux USA jusqu'à six enceintes Alexa ou Google Home par foyer, une par pièce. Le marché est énorme, ces machines pourraient nous accompagner au quotidien, pour surveiller notre santé, nous éduquer, nous aider et nous amuser, bref pour s'occuper de nous. Pour ces tâches, la machine est créée en vue d'être une sorte de compagnon numérique, assistant ou surveillant.

La communication avec les machines est avant tout un échange d'information avec trop souvent de l'« incommunication ». L'incommunication peut se produire même dans des circonstances où de l'information est communiquée, si l'information ne contient pas de message ou si le récepteur, par exemple le robot, ne peut pas décoder l'information contenue dans le message. Les machines sont loin d'avoir des capacités sémantiques suffisantes pour converser et partager des idées, mais elles pourront bientôt détecter notre malaise, notre stress, peut-être certains de nos mensonges.

L'empathie des humains pour les machines

La "*media equation*" de Reeves et Nass (1996) explique que nous appliquons les mêmes attentes sociales lorsque nous communiquons avec des entités artificielles, et que nous assignons inconsciemment à celles-ci des règles d'interaction sociale.

L'anthropomorphisme est l'attribution des caractéristiques comportementales ou morphologiques de vie humaine à des objets. Avec ce réflexe, à la fois inné et socialement renforcé, un objet qui semble être dans la douleur peut inspirer de l'empathie. Des études expérimentales ont montré la projection de réactions affectives et probablement empathiques envers des entités artificielles, par exemple des robots jouets volontairement endommagés.

Les chercheurs ont constaté que les humains ressentaient de l'empathie envers des robots maltraités, certes de moindre intensité qu'envers des humains maltraités, mais cette empathie n'existe pas envers des objets inanimés. Les recherches récentes, grâce à l'imagerie cérébrale, indiquent que les individus répondent de façon étonnamment semblable aux images émotionnelles des humains et à celles des entités artificielles. Si nous représentons les entités artificielles comme des humains, alors il n'est peut-être pas surprenant que nous réagissions avec émotion envers les agents artificiels comme nous réagissons envers les humains, cependant il n'est pas clair que des représentations de même complexité soient attribuées aux robots.

Les robots Nao, Pepper et Romeo sont des robots sociaux qui peuvent être aussi des miroirs de nos émotions, mais ils ne sont pas à proprement parler empathiques. Fan Hui, champion de Go qui a entraîné AlphaGo pour Google DeepMind, expliquait que jouer contre une machine, c'était un peu jouer contre soi-même, car on projetait ses émotions sur la machine qui les renvoyait comme un miroir. Le robot PARO le phoque, développé dès 1993 au Japon, a été commercialisé au Japon en 2005, puis aux États-Unis en 2009 (certification FDA en tant que robot thérapeutique).

Quand PARO est utilisé en EHPAD auprès de personnes âgées, ses réactions ne sont pas empathiques, elles ressemblent plus à des comportements expressifs d'animal de compagnie. Lors de nos tests en EHPAD (Garcia *et al.,* 2017) avec les robots Nao et Pepper qui détectaient les émotions et s'adaptaient en conséquence, les personnes âgées que nous avons rencontrées (une bonne cinquantaine de personnes de moyenne d'âge de 85 ans), qui n'étaient pas sous tutelle, ont eu des réactions de curiosité et d'amusement. Elles n'ont cependant pas considéré que le robot les « comprenait », mais qu'il pouvait détecter certains de leurs comportements.

Un enfant de deux ans sait que son doudou n'est pas vivant, et pourtant il lui parle comme si celui-ci l'était. Les psychologues parlent des poupées et doudous comme d'« objets transitionnels ». L'objet transitionnel est donc un objet privilégié, choisi par l'enfant, généralement doux au toucher. Il permet au bébé de lutter contre l'angoisse, il est la première possession. Il n'est perçu ni comme faisant partie de la mère, ni comme étant un objet intérieur. Il permet le cheminement de l'enfant du subjectif vers l'objectif. PARO le robot en peluche peut être vu comme un objet transitionnel. Les robots sociaux sont également avant tout des objets techniques qui enregistrent nos données pour les transférer par exemple à un médecin.

Les robots assistants de vie s'insèrent aussi dans un écosystème qui comprend de nombreux acteurs : la famille, les aides-soignants, les médecins.

Les troubles de comportements face aux robots sont étudiés en psychologie et en psychiatrie, notamment par Serge Tisseron, psychiatre (Tisseron, 2015). Kate Darling, chercheuse au Media lab du MIT, étudie les réactions empathiques des personnes devant des robots, notamment en cas de maltraitance des robots, avec l'idée qu'il faudrait accorder aux robots une protection juridique comme on l'a fait pour les animaux (Darling, 2016). Pourtant, les animaux sont des organismes

vivants qui peuvent ressentir des émotions et souffrir, ce qui n'est absolument pas le cas des robots qui ne font que simuler des émotions.

Les émotions des machines

Donner aux machines des capacités d'interprétation et de simulation émotionnelle est indispensable pour construire des systèmes capables d'interagir socialement et de mieux communiquer avec les humains. Les applications dans le but d'aider les personnes dépendantes, dans le grand âge ou pour différentes pathologies dégénérescentes sont nombreuses. La sphère des émotions que l'on pouvait penser propre à l'humain envahit les machines, qui se rapprochent des capacités humaines. Dans certains cas, une hésitation, un souffle de la machine donne l'impression qu'elle est en « vie ». Donal Davinson, philosophe américain, décrit le monisme anomique comme « l'union exprimée par deux langages différents sans traduction ». Il n'y a pas de raison causale entre l'esprit et le corps. Plus la machine a l'air fragile, plus on peut l'humaniser et être ému devant elle, même si à proprement parler elle n'est pas empathique.

Le philosophe Spinoza et plus particulièrement son ouvrage *L'Éthique* (1677) est une source d'inspiration pour expliquer le monde d'aujourd'hui et les relations entre le corps et l'esprit. Spinoza explique que l'organisme se fabrique lui-même. Les affects et les haines, lieux par excellence où sont unis le corps et l'esprit, sont sources d'aliénation si nous les subissons et/ou de liberté si nous en comprenons les mécanismes sous-jacents. Grâce à l'exploration du cerveau, nous pouvons aujourd'hui prouver les déclarations du philosophe qui étaient contraires au sens commun à son époque. Ainsi, les expressions corporelles précèdent le sentiment. Corps et esprit sont mélangés, nous apprend Spinoza. Mais les machines n'ont pour l'instant pas de corps au sens de viscères, d'hormones, de peau ; elles n'ont pas d'intention, de plaisir et de désir propres, c'est-à-dire pas de « *conatus* » au sens entendu par Spinoza. Le vivant peut par conséquent être défini comme autonome et ayant la possibilité de se reproduire.

À l'heure actuelle, la relative autonomie des robots est toujours programmée par l'humain. La faculté d'apprentissage programmée peut offrir plus ou moins de liberté à la machine. Donner à un robot la capacité d'apprendre seul, en interaction avec l'environnement et les humains, est le Graal des chercheurs en intelligence artificielle. Si les robots apprennent seuls, il sera souhaitable de leur enseigner les valeurs communes et morales de la vie en société. Cette faculté constitue cependant une rupture technologique et juridique, et soulève de nombreuses questions éthiques. Ces robots peuvent être, d'une certaine manière, créatifs et autonomes dans leurs prises de décision, si on les programme pour cela.

Vouloir recopier l'intelligence de l'homme sur une machine est très narcissique, car que connait-on de notre intelligence ? Nous ne connaissons pas le substrat de la pensée et n'avons pas conscience que certains de nos organes sont autonomes, nous ne sommes conscients que d'une petite partie de nos perceptions et de notre activité cérébrale. Il n'existe pas de terme plus polysémique et sujet à interprétation que celui de « conscience » : il évoque pour certains la conscience de soi, pour d'autres la conscience du prochain, ou encore la conscience phénoménale, la conscience morale, etc.

Avec une conception philosophique matérialiste de la vie, on peut considérer que l'ordinateur et le cerveau humain sont des systèmes comparables, capables de manipuler des informations. Les modélisations numériques les plus performantes, comme le *deep learning* (apprentissage profond), s'appuient sur une modélisation simplifiée du neurone (neurone formel), intégrée dans une machine à états discrets simulée sur ordinateur. Le nombre de couches cachées de l'architecture du modèle correspond à la profondeur. Pour l'instant, nous sommes très loin de la complexité du vivant !

Les systèmes actuels d'intelligence artificielle ont la capacité de calculer des corrélations de faits, avec les approches d'apprentissage profond par exemple, de prendre des décisions et d'apprendre, mais sans en avoir conscience. Certains prototypes de robots ont pourtant déjà des embryons de niveau de « conscience » comparables à ceux que décrit Stanislas Dehaene. Ils sont simulés par des mécanismes de partage de connaissances et d'introspection. Pour autant, ces machines ne sont pas conscientes comme peut l'être un humain, elles n'ont ni conscience morale ni conscience phénoménale associée à une expérience qualitative telle que la sensation de chaud ou de froid, le sentiment d'anxiété, ..., car elles n'ont pas de viscères ni de ressenti, à moins, là encore, de les simuler.

Les premiers robots apprenants et communicants, dotés de capteurs de douleur et de plaisir (Asada, 2015), interagissent par des procédés simples : des capteurs sur leur corps, une caméra et un microphone qui leur permettent d'associer un visage et une voix aux signes expressifs qu'ils reçoivent. Leurs indicateurs d'expression (sons émis, diodes) permettent aux humains de comprendre leurs états. Pour produire des effets rapides, il faut des actions physiques : une caresse sur la tête, et ils associent un critère positif à la personne qu'ils voient ; une tape sur la tête, et ils lui associent un critère négatif. Ces machines apprennent et adaptent leur comportement aux contextes dans lesquels ils se trouvent. Mais les interprétations sont encore très limitées. Par exemple, un robot assistant d'étudiants en chirurgie dentaire devrait informer au mieux sur les grimaces de douleur, et anticiper des gestes afin d'alerter sur la proximité d'un nerf et une possible souffrance. Faut-il que les robots s'approchent le plus possible des humains ? Une conscience artificielle, dotée de sentiments, de pensées et de libre arbitre sans programmation humaine a encore peu de chance d'émerger spontanément avec les architectures actuelles d'ordinateurs.

Conclusion

La vie au quotidien avec des robots pourrait entraîner des risques sociaux à long terme qu'il faut anticiper pour tirer bénéfice de ces machines. L'addiction et l'isolement, ainsi que le report d'autonomie sur la machine, la confusion entre la machine et l'humain sont des déviances dont il faut se préoccuper. Un des risques, particulièrement pour les personnes fragiles, est d'oublier qu'un robot est connecté et programmé. La capacité d'un robot de s'adapter à son propriétaire humain pourrait bien être utilisé pour lui faire accomplir certains choix plutôt que d'autres, notamment pour l'aider à mieux gérer les déviances de comportements de ce dernier (pathologie sexuelle, addiction à la drogue), mais aussi pour des causes moins louables dans le domaine de la consommation. Un autre risque est d'oublier qu'un robot ne ressent rien, n'a pas d'émotions, n'a pas de conscience et n'est pas vivant. Il est possible de ressentir de l'empathie pour un robot et de parler de souffrance pour un robot. Il est important que les personnes âgées, qui peuvent mettre leur vie en danger pour venir en aide à leur robot, se rendent compte qu'un robot ne souffre pas même s'il tombe, il faut qu'elles soient conscientes que ce n'est qu'un objet programmé.

Les robots empathiques soulèvent de nombreuses questions éthiques, juridiques et sociales. Ces questions prégnantes ne sont évoquées que depuis peu. 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 souhaitons en faire. 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 sur le long terme de la « coadaptation » humain-machine (la machine s'adaptera à l'humain et l'humain à la machine).

Le contrôle par des humains sera toujours primordial. Il est nécessaire de développer des cadres éthiques pour les robots sociaux, notamment dans le domaine de la santé, et de comprendre le

niveau de complémentarité humain-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.

Bibliographie

ASSADA M. (2019), "Artificial Pain: empathy, morality, and ethics as a developmental process of consciousness", *Philosophies* 2019, 4, 38; doi: 10.3390.

BOSTROM N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press.

DAMASIO D. (1994), *L'Erreur de Descartes*, Éditeur Odile Jacob.

DARLING K. (2016), « Faut-il accorder une protection juridique aux robots de compagnie ? », dans le livre d'Alain Bensoussan, Yannis Constantinides, Kate Darling, Jean-Gabriel Ganascia et Olivier Tesquet, *En compagnie des robots,* Éditeur Premier parallèle.

DEHAENE S. (2014), *Le code de la conscience,* Éditeur Odile Jacob.

DEVILLERS L. (2017), *Des robots et des hommes : mythes, fantasmes et réalité,* Éditeur Plon.

DUMOUCHEL P. & DAMIANO L. (2016), *Vivre avec les robots*, Éditeur Seuil.

GARCIA M., BECHADE L., DUBUISSON DUPLESSIS G., PITTARO G. & DEVILLERS L. (2017), "Towards Metrics of Evaluation of Pepper Robot as a Social Companion for Elderly People", International Workshop on Spoken Dialogue Systems (IWSDS), 8 p.

PICARD R. (1997), *Affective computing*, MIT Press.

SPINOZA B. (1677), *L'Éthique,* Livre III, Proposition II, et SCOLIE, (Point-Essais), trad. Bernard Pautrat, pp. 207-209.

REEVES B. & NASS C. (1996), *The Media Equation,* CSLI Publications, Stanford University.

TISSERON S. (2015), *Le jour où mon robot m'aimera : vers l'empathie artificielle,* Éditeur Albin Michel.