Publié dans "Revue de Philosophie
et Psychology 11, pages 881-897
(2020)" qui doivent être cités pour faire
référence à ce travail.

Comprendre l'IA - Pouvons-nous et devons-nous faire preuve d'empathie avec des robots ?

Susanne Schmetkamp1

Published online: 28 April 2020

Abstrait
En élargissant le débat sur l'empathie envers les êtres humains, les animaux ou les personnages fictifs pour inclure les relations homme-robot, cet article propose deux perspectives différentes pour évaluer la portée et les limites de l'empathie envers les robots : la première est
épistémologique, tandis que la seconde est normative. L'approche épistémologique nous aide pour clarifier si nous pouvons sympathiser avec l'intelligence artificielle ou, plus précisément, avec robots sociaux. L'énigme principale ici concerne, entre autres choses, de quoi il s'agit exactement avec lesquels nous sympathisons si les robots n'ont pas d'émotions ou de croyances, puisqu'ils n'ont pas une conscience dans un sens élaboré. Cependant, en comparant les robots avec des robots fictifs personnages, le document montre que nous pouvons encore sympathiser avec les robots et que bon nombre des Les récits existants sur l'empathie et la lecture mentale sont compatibles avec une telle vision. Par ainsi Ce faisant, l'article se concentre sur l'importance de la prise de perspective et affirme que nous attribuent également aux robots quelque chose comme une expérience de perspective. L'approche normative examine l'impact moral de l'empathie avec les robots. À cet égard, le document discute de manière critique de trois réponses possibles : stratégique, anti-barbarisation et pragmatiste. Cette dernière position est défendue en soulignant que nous sommes de plus en plus contraints d'interagir avec les robots dans un monde partagé et d'intégrer les robots dans notre morale la considération doit être considérée comme une partie intégrante de notre compréhension de soi et des autres.

Mots-clés Empathie. Intelligence artificielle. Des robots humanoïdes. Interaction. Prise de perspective. Personnages de fiction. Éthique

1. Introduction

Les débats sur l'empathie ou, plus largement, sur la compréhension interpersonnelle ont été un pilier de l'érudition dans un large éventail de disciplines. Cependant, même si beaucoup de choses ont a été écrit sur la capacité humaine à sympathiser avec des personnes réelles ou des personnages fictifs

*Suzanne Schmetkamp
Susanne.schmetkamp@icloud.com

1 Département de philosophie, Université de Fribourg, Fribourg, Suisse

(pour des aperçus récents, voir Coplan et Goldie 2011 ; Maibom 2017), jusqu'à récemment, les philosophes ont quelque peu négligé le rôle de l'empathie dans les interactions homme-robot (HRI) (cf. Brinck et Balkenius 2018 ; Lin et al. 2017). Pourtant, parallèlement au nombre croissant d'études sur les émotions ou d'autres caractéristiques des systèmes d'intelligence artificielle1, la possibilité et la nécessité d'interagir et de sympathiser avec différentes formes d'intelligence artificielle ont suscité un grand intérêt philosophique, en particulier avec ce que l'on appelle les systèmes sociaux. robots.2 Cet intérêt a également donné lieu à des discussions sur la valeur de l'empathie pour la société en général ou pour les soins de santé et la thérapie en particulier (Coeckelbergh 2018 ; Darling 2016 ; Engelen 2018 ; Loh 2019 ; Misselhorn sous presse ; Vallor 2011). Il devient clair qu'à l'avenir, les robots et les androïdes – c'est-à-dire des robots qui ressemblent à des humains – deviendront des acteurs plus ou moins indépendants dotés de compétences sociales. À ce titre, ils sont amenés à devenir des compagnons importants et de plus en plus capables d'établir des relations avec les êtres humains (Benford et Malartre 2007 ; Breazeal 2002 ; Dumouchel et Damiano 2017). De plus, les systèmes d'apprentissage profond (Kasparov 2017) seront déployés dans de nombreuses (jusqu'à présent) professions humaines, ce qui non seulement améliorera ou facilitera certaines tâches ou défis (dans la recherche médicale, par exemple) ; Ils pourraient également nous obliger à reconsidérer certains concepts clés tels que l'intelligence, l'action, la conscience, l'autonomie, les émotions ou les perspectives (Schneider sous

Comme des études l'ont montré (Leite et al. 2013), la forme et le succès des relations homme-robot dépendent souvent de caractéristiques humaines, telles que la capacité des robots à exprimer des émotions, à interagir et à exécuter des décisions (plus ou moins) autonomes. . Ces capacités sont également importantes pour la compréhension empathique réciproque.3 Alors que les êtres humains reconnaissent et attribuent également des émotions par rapport à des formes virtuelles abstraites ou même à des dispositifs techniques (les meilleurs exemples étant les smartphones et les ordinateurs), pour notre interaction coopérative et collaborative avec les robots. – en particulier dans le contexte médical ou des soins de santé – une forte ressemblance humaine pourrait être cruciale pour la réussite de ces interactions. À mesure que la présence de robots sociaux humanoïdes dans la société augmente, la nécessité d'examiner et de façonner nos interactions avec eux augmente également. La génération actuelle de robots est déjà capable d'exprimer toute une série d'émotions : l'IA humanoïde "Sophia", par exemple, connaît 60 expressions faciales différentes et semble même capable de communiquer avec humour et ironie. Cependant, les robots n'ont pas de conscience au sens d'expérience subjective4 et ne possèdent pas d'humour ou d'émotions au sens élaboré (Boden 2016 ; MacLennan 2014 ; Scheutz 2011). Pourtant, ils pourraient avoir quelque chose qui peut être considéré comme analogue aux émotions humaines et à certains processus mentaux. De plus, à la lumière des récentes découvertes de la philosophie de la cognition incarnée, il se pourrait que le corps et le comportement humains ainsi que la cognition « étendue » soient ce qui aide les humains à reconnaître les androïdes comme des partenaires et comme semblables à eux à certains égards, tandis que res

[1] Un groupe du MIT Media Lab et de l'association de normalisation IEEE plaide en faveur du concept d'intelligence « étendue » plutôt que « artificielle ». Grâce à ce nouveau récit « étendu », ils veulent garantir que les robots ne se substituent pas aux êtres humains, mais plutôt les soutiennent et coopèrent avec eux. Ensemble, ils ont créé le Council on Extended Intelligence CXI, voir https://globalcxi.org (dernière consultation le 12.12.2019).

[2] Un projet financé par l'ERC, situé à l'Université de Glasgow et dirigé par Emily Cross, examine en particulier la socialisation des êtres humains grâce à l'intelligence artificielle et l'importance de l'interaction et des relations avec les robots pour la cognition sociale. L'accent est mis sur la capacité des robots à être des compagnons, http://www.so-bots.com (dernière consultation le 20.12.2019).

[3] Concernant le phénomène de la « vallée étrange », voir ci-dessous.

[4] Du moins quand on suit une position anti-physicaliste.

chez les autres (Benford et Malartre 2007, 181 ; Hoffmann et Pfeifer 2018 ; Newen et al. 2018).5

L'empathie est largement considérée comme un moyen crucial d'appréhender et de revivre les états mentaux des autres par la lecture de pensées, le partage émotionnel et/ ou ou une co-expérience expérientielle (voir par exemple Engelen et Röttger-Rössler 2012 ; Goldman 2006 ; Stueber 2018 ; Zahavi 2014).6 En philosophie, l'empathie se distingue généralement de la contagion affective et de la sympathie ou compassion morale.7 Alors que cette dernière vise la bien-être des autres et veut le promouvoir (ou du moins ne pas l'entraver) (Darwall 1998), l'empathie conduit en premier lieu à la compréhension des processus mentaux des autres – tels que les émotions ou les croyances. Contrairement à la simple contagion émotionnelle, une différenciation soi-autre doit être mise en place (De Vignemont et Jacob 2012). Il y a toujours un débat important sur ce point et diverses définitions et approches ont été proposées qui cherchent à répondre à des questions telles que : Comment percevons-nous et accédons-nous aux états et aux expériences des autres ? Comment caractériser le processus empathique Quel est le résultat de ce processus ? D'une manière générale, les théories prédominantes – issues de la philosophie de l'esprit ou de la phénoménologie – sont la Théorie des Neurones Miroirs ou Théorie de la Résonance (MNT) (Gallese 2001), la Théorie Théorie (TT) (Fodor 1987 ; Gopnik et Wellman 1994), la Théorie de la Simulation (ST) (De Vignemont et Jacob 2012 ; Goldman 2006, 2011 ; Stueber 2006), la théorie de la perception directe (DPT) (Zahavi 2011) avec ses variantes de la théorie de l'interaction (IT) (Gallagher 2008, 2017) et de la théorie de la narration (NT) (Gallagher et Hutto 2008). En outre, il existe des théories hybrides et pluralistes qui combinent deux ou plusieurs approches, telles que la perception directe et l'imagination8 (Schmetkamp 2017, 2019 ; Dullstein 2013 ; pour d'excellents aperçus, voir Newen 2015 ; Stueber 2018 ; Zahavi 2014 ; 2018). Compte tenu de cet ensemble diversifié d'approches, il convient d'établir d'autres distinctions entre l'empathie cognitive (telle que TT) et l'empathie affective (telle que ST ou MNT).9

En se demandant si nous pouvons réellement sympathiser avec les robots, la section 2 se concentrera sur les nombreuses approches épistémiques. dimensions des relations empathiques avec les robots : Que percevons-nous et comprenons-nous s'il n'y a pas réellement d'émotions,

---

[5] L'article se concentre principalement sur les robots humanoïdes. L'une des raisons à cela est que cela contribue à limiter la portée du document ; Une autre raison est l'hypothèse selon laquelle les caractéristiques humaines facilitent effectivement notre interaction sociale avec l'intelligence artificielle et rendent plus plausible le fait que nous traitions les robots comme des partenaires sociaux. Cependant, nous pouvons également sympathiser avec des formes plus abstraites d'IA en leur attribuant des états émotionnels et des motivations (voir Isik, Koldeewyn, Beeler et Kanwisher 2017). Je suis très reconnaissant à un critique pour cette remarque.

[6] Il est très controversé de savoir si l'empathie présuppose ou implique une réflexion affective, une lecture d'esprit théorique, une prise de perspective simulée, une compréhension émotionnelle et/ou une compréhension expérientielle, et il n'y a actuellement aucune fin en vue à ce débat (voir par exemple Zahavi 2018). De nombreux philosophes soulignent que la lecture de pensées est quelque chose de distinct de l'empathie, et que l'empathie est « quelque chose en plus ». Ici, cependant, j'ai essayé d'appliquer toutes les différentes approches. Mais ma propre position est phénoménologique.

[7] L'un des problèmes de tout le débat est cependant qu'il n'existe pas de consensus conceptuel sur ce qu'est et ce qu'implique l'empathie. Le projet financé par l'ERC sur les robots sociaux, par exemple, définit l'empathie comme impliquant à la fois une correspondance émotionnelle et un comportement prosocial. En philosophie, cependant, l'empathie n'est généralement pas considérée comme une émotion ou une attitude morale (voir Cross et al. 2018 ; Zahavi 2018).

[8] Par exemple, en se référant aux positions classiques de Stein ou de Dilthey et en combinant perception directe et re-présentation imaginative (« Vergegenwärtigung ») (voir aussi Gallagher 2019).

[9] Kanske (2018) fait la distinction entre l'empathie affective proprement dite et la théorie cognitive de l'esprit. Alors que la première capacité nous permettrait de ressentir ce que ressentent les autres, l'autre nous aiderait à comprendre ce que pensent ou croient les autres. Bien que je reconnaisse les différences, je ne distinguerai pas ici la mentalisation de l'empathie, mais j'examinerai différentes formes de compréhension des autres esprits sous le terme générique d'empathie puisque c'est le terme central du débat philosophique actuel.

perspectives dans un sens riche ? Ou les robots ont-ils quelque chose de similaire aux émotions, aux croyances et aux expériences ? Ont-ils une vision individuelle du monde (Schmetkamp 2017) ou un récit (Gallagher 2012), puisqu'au moins ils sont incarnés et contextualisés ? En comparant les robots avec des personnages fictifs, la réponse sera affirmative : oui, dans une certaine mesure, nous pouvons sympathiser avec les robots de manière cognitive, affective et même expérimentale, soit en déduisant, en ressentant, en interagissant ou en imaginant comment ils perçoivent et se déplacent. leur monde, tout comme nous comprenons de manière plurielle (Vaage 2010) comment un personnage de fiction (par exemple dans un film) perçoit son monde, agit et ressent. L'aspect crucial ici sera que nous attribuons une perspective individuelle à l'autre. Nous le comprenons indépendamment du fait que cette perspective soit seulement racontée, projetée ou programmée.10 La deuxième question, qui sera discutée dans la section 3, demande si nous devons sympathiser avec les robots. Il y a deux aspects à cette question : soit nous pouvons nous demander si l'empathie envers les robots a une simple fonction stratégique en ce qui concerne l'amélioration de la compréhension réciproque dans la relation homme-robot, soit nous pouvons nous demander si l'empathie a un impact éthique tel que nous ont le devoir de sympathiser avec les robots (pour un aperçu sur le thème de l'éthique et de l'IA, voir Boddington et al. 2017). Si, par exemple, nous pouvons comprendre épistémiquement ce que les robots perçoivent, envisagent ou pourraient même « ressentir », nous sommes également capables de prédire ce qu'ils feront ensuite. En général, cela pourrait être utile en termes de nos interactions avec eux.11 Il s'agit clairement d'un « devrait » stratégique ou rationnel. Le deuxième sens de la question donne lieu à une réponse normative : devons-nous faire preuve d'empathie envers les autres au sens moral ? Et qu'est-ce que, d'un point de vue moral, est-ce qu'eux ou nous – en tant qu'empathiques – gagnons en cela ? Considérant cette question, à première vue, une réponse kantienne pourrait être évidente, qui suit le précédent établi par la vision de Kant sur les animaux et qui peut être modifiée pour s'appliquer à l'intelligence artificielle : à savoir que nous devrions faire preuve d'empathie, selon l'argument, afin d'éviter. » « barbarisation morale ». En fin de compte, l'article n'empruntera ni la voie stratégique ni la voie kantienne, mais proposera plutôt une réponse pragmatiste et relationnelle. Cette réponse est liée aux deux autres. Cependant, il souligne l'impact de l'interaction et de la compréhension de soi-même.

2 Pouvons-nous sympathiser avec les robots ?

Pour des raisons d'espace, je me concentrerai sur les robots qui ont à la fois un visage et un corps, affichent des expressions et des comportements humains, sont destinés à interagir avec des êtres humains et sont donc incarnés et intégrés dans notre vie quotidienne et, en tant que tels, sont soumis aux contraintes sociales. évaluation par les humains. Une deuxième raison de cette focalisation est l'hypothèse selon laquelle les robots dotés de caractéristiques et d'expressions humaines sont probablement encore plus capables de développer la confiance et de susciter des réponses émotionnelles similaires à celles des vrais humains (Brinck et Balkenius 2018 ; Mori 2005) et sont, à cet égard , plus susceptibles ... être reconnus et acceptés comme partenaires dans l'interaction sociale. Bien que des études en psychologie cognitive aient montré que nous pouvons également faire preuve d'empathie ou lire dans les pensées de systèmes qui ont peu de ressemblance physique (Bretan et al. 2015), une appa

[dix] L'article se concentre sur la question épistémologique. Cela ne répondra pas à la question métaphysique de savoir si les robots ou l'IA ont une conscience.

[11] Concernant le déploiement de systèmes de Deep Learning en médecine, entre autres fonctionnalités, il est nécessaire de faire confiance à la machine intelligente et de comprendre ce qu'elle va faire, par exemple dans une interaction robot médical-patient.

important pour l'utilisation de robots comme soignants ou collègues dans le domaine de la santé (Vallor 2011).12 Mais quel type d'empathie est ici en jeu ? Reproduisons-nous les expressions des robots ? Devons-nous interpréter et prédire leur comportement ? Ou bien faisons-nous preuve d'empathie d'une manière plus phénoménologique et interactive ?

De manière très minimale, l'empathie peut être définie comme la capacité humaine à comprendre les états mentaux des autres et à les revivre d'une manière ou d'une autre, même si la question de savoir si le sujet empathique a besoin de ressentir la même chose que l'autre reste un sujet de débat. Certaines théories limitent les objets de l'empathie aux émotions des personnes et à leurs expressions comme indicateurs d'états affectifs. D'autres sont plus larges et incluent d'autres processus cognitifs comme objets d'empathie, tels que les croyances, les désirs et leurs raisons respectives (pour un aperçu, voir Batson 2009 ; Slote 2017). Une définition importante implique une condition d'isomorphisme : l'empathique et la cible sont dans le même ou au moins un état affectif similaire (De Vignemont et Singer 2006). Cependant, comme certains critiques l'ont soutenu, l'empathie n'implique pas nécessairement que nous reproduisons l'état mental des autres (Zahavi et Michael 2018). Nous n'avons pas non plus à nous soucier de l'autre dans un sens plus élaboré.

Comme chacun le sait, le « battage médiatique » actuel autour du thème de l'empathie peut être largement attribué à la découverte des « neurones miroirs » (Iacoboni et al. 1999 ; 2011).
D'une manière générale, les neurones miroirs sont les neurones situés dans une zone du cerveau qui sont déchargés à la fois pour l'observation et l'exécution d'actions similaires. Ce processus d'imitation a été appliqué à la compréhension des émotions humaines : en observant l'expression affective d'une autre personne - comme un visage triste - les mêmes neurones seraient comme si nous - en tant qu'observateurs - avions fait une grimace et ressentions nous-mêmes de la tristesse. Alors que cette théorie a été largement critiquée (Hickok 2014) et rejetée en tant que théorie de l'empathie, d'autres l'ont invoquée dans leur approche plus élaborée de l'empathie. Dans son exposé sur la théorie de la simulation (ST), Alvin Goldman, par exemple, fait la distinction entre une forme de lecture mentale de bas niveau et une forme de haut niveau de lecture mentale ou une « voie miroir » et une « voie reconstructive », bien que la « résonance » émotionnelle soit mise en œuvre. via les deux routes (Goldman 2006, 2011). Les neurones miroirs constituent la partie principale des processus de bas niveau par lesquels nous comprenons immédiatement et automatiquement les états mentaux d'une autre personne. À un niveau plus complexe et plus élevé, nous simulons l'état de l'autre dans notre propre esprit, puis arrivons à la connaissance de ce que ressent l'autre, non pas en diffusant une théorie, mais plutôt en imitant le comportement des autres dans notre esprit et en projetant ensuite notre propre processus mental sur l'autre. Selon ST, nous simulons, via une perspective à la première personne, la situation de l'autre et utilisons nos propres mécanismes mentaux pour générer des pensées, des croyances, des désirs et des émotions. Au cours des deux dernières décennies, ST - aux côtés de son adversaire Theory Theory (TT) - a dominé le débat sur la lecture mentale. TT affirme que notre compréhension des autres esprits repose essentiellement sur la psychologie populaire, qui est soit innée, soit acquise pendant la petite enfance (Baron-Cohen 1995).
TT suppose que nous faisons des inférences fondées sur la théorie afin de comprendre les autres … et leurs désirs. Le TT a été critiqué pour être trop théorique et trop général (Zahavi 2014 ; mais cf. Fodor 1987). Ses détracteurs affirment que TT ne prend pas en compte les autres bétons et ne

[12] Cependant, empiriquement, il reste incertain si les robots doivent effectivement ressembler à des humains en HRI (Brinck et Balkenius 2018).

[13] Un problème, bien sûr, est de savoir comment nous comprenons le terme « compréhension ». Monika Dullstein (2012) a ont montré que les récits de la Théorie de l'Esprit utilisent une notion assez différente des récits phénoménologiques.

reconnaître l'incarnation et l'intégration des autres. De plus, TT et ST semblent être dupes d'une fausse vision cartésienne occlusionniste de l'esprit, comme si nous ne pouvions pas percevoir ce qui se passe dans l'esprit d'autrui (Zahavi 2011, 2014) . En revanche, les récits phénoménologiques mettent l'accent sur l'incarnation et l'ancrage de l'être humain et soutiennent que nous sommes capables de voir directement dans le visage et les expressions corporelles de l'autre ce qu'il vit : dans cette perspective, nous n'avons pas besoin de déduire ou d'imaginer ce qu'il ressent. ; il suffit de le percevoir. De plus, nous le faisons dans un contexte situationnel partagé et par interaction. Pour cette raison, une telle approche est appelée théorie de la perception directe (DPT).
(Zahavi 2011, 2014) ou théorie de l'interaction (TI) (Gallagher 2001; 2012). Contrairement au TT, le DPT et l'IT soutiennent que nous n'adoptons pas une position de troisième personne envers les autres et ne les observons pas. En outre, DPT et IT soutiennent également que nous n'avons pas d'accès indirect et imaginatif aux autres. Au lieu de cela, nous interagissons socialement d'une manière seconde personnelle, par laquelle deux « vous » se reconnaissent de manière complémentaire et réciproque (Dullstein 2012 ; Engelen 2018 ; Zahavi et Michael 2018). Les limites du DPT apparaissent évidemment dans des situations où l'autre n'est pas présent à nous : par exemple, lorsque quelqu'un nous raconte l'histoire de quelqu'un d'autre, ou si nous lisons un roman, regardons un film ou voyons une pièce de théâtre où les expériences des autres sont mises en avant. d'une manière ou d'une autre, médiatisés par quelqu'un d'autre (par exemple un narrateur), nous n'avons pas de rencontres directes. Ainsi, tous ces cas sont des cas où l'autre est donné par la narration, parfois même dans un cadre fictionnel. C'est pourquoi certains philosophes ajoutent qu'un récit est essentiel pour comprendre les autres esprits ou susciter de l'empathie au-delà du sens le plus éléme
Daniel D. Hutto (2008) a formulé l'hypothèse de la pratique narrative (NPH). Selon cette thèse, nous ne comprenons les raisons d'agir des autres, leurs croyances et leurs désirs que lorsque nous prenons également en compte les circonstances individuelles, l'histoire du sujet, sa situation actuelle, ses espoirs et ses expériences, ses traits de caractère, etc. Autrement dit, selon NPH, pour appréhender la situation d'une personne, il faut s'appuyer sur « son histoire » (Gallagher 2012). Cette vision permet également de sympathiser avec « les monstres ou les extraterrestres venus d'autres planètes, tels que décrits dans les films » (Gallagher 2012). Cependant, une sorte d'imagination est ici nécessaire : il existe de nombreux cas - non seulement mais surtout dans notre relation avec la fiction - où nous comptons sur notre imagination comme moyen de rendre disponible quelque chose qui ne nous est pas présent. Même l'une des premières pionnières des approches phénoménologiques de l'empathie, à savoir Edith Stein (1989), affirmait que l'imagination ou la « re-présentation » mettait en scène un [14] joue un rôle crucial au sein d'un multi processus de compréhension empathique. C'est pourquoi certaines théories de l'empathie combinent une approche seconde personnelle avec une forme de représentation imaginative de la situation concrète, du récit et/ou de la perspective de l'autre (Schmetkamp 2019 ; Gallagher et Gallagher 2019).15 Peu importe si nous

devrions ou non considérons tous ces différents récits comme des théories de l'empathie ou plus largement comme des théories de la compréhension interpersonnelle, pour chaque récit nous pouvons nous poser la question suivante d'un point de vue descriptif et épistémologique : Comment pouvons-nous

[14] Il est difficile de donner une traduction exacte du concept de « Vergegenwärtigung » de Stein. La traduction anglaise (Stein 1989) utilise la « représentation » ou « l'acte représentationnel » (Stein 1989 : 8) comme une « donation » représentée non primordiale des expériences indirectes ou des autres (analogue à la mémoire, à l'attente et au fantasme) (ibid.). Dans le débat, on oublie souvent que Stein propose un modèle d'empathie par étapes, selon lequel le premier niveau est la perception directe de l'expérience de l'autre, le deuxième niveau étant une sorte de réflexion et de prise de perspective (Stein 1989 : 10) .

[15] Gallagher a récemment défini l'empathie comme suit : « L'empathie pourrait [...] non seulement [compter] comme quelque chose qui se produit, mais aussi comme une méthode ; et cela […] implique de se mettre dans la perspective ou la situation de l'autre. (2018). Ce faisant, Gallagher a élargi son approche narrative vers une approche perspective (combinant le récit avec la perspective subjective).

sympathiser avec l'IA, par exemple les robots ressemblant à des humains, si le récit respectif était le plus plausible ? Par exemple, si nous observons les expressions et/ou les actions d'un robot, on pourrait affirmer que nous résonnons et imitons automatiquement son comportement expressif. Si nous voulons prédire ce que le robot fera ensuite, nous pourrions également déployer une théorie psychologique populaire et en déduire les raisons pour lesquelles il agit. Nous pourrions simuler ce que nous ferions si nous étions dans leur situation, puis projeter notre expérience sur eux. Ou encore, lors de rencontres directes, nous pourrions être capables de percevoir leurs actions de manière interactive. Nous pourrions considérer leur ancrage dans un contexte narratif et comprendre la structure intentionnelle de leurs émotions sans en même temps reproduire leur contenu « qualitatif ». D'un point de vue empirique, ces modes de compréhension interactifs existent certainement.

Cependant, certaines objections métaphysiques et épistémologiques évidentes peuvent être soulevées. Le principal problème est que les robots ne ressentent et n'éprouvent rien. Ils n'ont pas non plus réellement d'états mentaux tels que des désirs ou des croyances, car ils n'ont aucune conscience. Cela dit, il semble également étrange de parler du point de vue individuel ou du récit personnel d'un robot. Dans la mesure où l'empathie s'adresse aux états mentaux et à « l'être au monde » de quelqu'un, la réponse serait : nous ne pouvons pas sympathiser avec les robots.

Pourtant, deux réponses possibles pourraient être données : premièrement, les « états mentaux » des robots sont souvent décrits comme des « états informatiques » considérés comme ayant une structure analogue aux états mentaux humains. Donc, si nous supposons que les robots ont quelque chose qui est comparable aux états mentaux humains, ont-ils aussi quelque chose comme des émotions ou des expériences avec lesquelles nous sympathisons ? Selon certaines conceptions philosophiques actuelles des émotions, les états ou processus émotionnels présentent une structure complexe composée de composantes cognitives et affectives (De Sousa 1987 ; Nussbaum 2001) : lorsque nous ressentons de la colère, notre colère est dirigée vers un objet que nous évaluons comme ennuyeux. . En psychologie, cela est également appelé théorie de l'évaluation, ce qui implique que nous portons des jugements sur les objets de notre environnement en fonction de leur pertinence par rapport à nos objectifs. Si les émotions n'étaient constituées que de cette simple composante cognitive, nous pourrions supposer que les robots ont des émotions au sens m Les robots, pourrions-nous dire, agissent pour un ensemble de raisons basées sur un ensemble de croyances sur le monde. Cependant, les émotions peuvent englober bien plus que cela : la colère, par exemple, est également ressentie au niveau sensationnel et corporel ; cela semble, par exemple, frustrant et rétrécissant. Cela dit, la colère a aussi des connotations négatives, dont nous prenons conscience de manière proprioceptive (Colombetti 2013). Cependant, s'ils ne sont pas purement virtuels, les corps des robots sont constitués de métal ou de plastique et, plus important encore, ils ne sont pas liés à un riche concept de conscience : dans le sens où ils se perçoivent comme un être émotionnel. Il ne peut pas ressentir de manière auto-référentielle ce que signifie être dans un corps en plastique. De plus, comme l'ont soutenu les récits narratifs d'émotions, les émotions complexes sont généralement intégrées dans un cadre narratif : nous pouvons raconter une histoire sur leur excitation et leur développement (Goldie 2000). Enfin et surtout, les êtres humains sont capables de gérer de manière créative leurs sentiments et leurs émotions : ils peuvent apprendre de nouvelles émotions et ils sont capables d'en modifier certaines et d'en cultiver d'autres.

Mais cela pourrait aussi être possible pour et avec les robots. Le point crucial ici est que nous attribuons également, de manière parfaitement intuitive, des émotions aux machines. Lorsque nous collaborons avec des robots, nous pouvons adopter une « position intentionnelle ». Ce concept, issu des travaux de Daniel Dennett, implique que l'on traite un objet dont on veut prédire le comportement comme un agent rationnel ; nous attribuons des croyances et des désirs et, sur cette base, nous prédisons son comportement (Dennett 1987). Mais cette approche repose néanmoins sur une théorie de la lecture mentale et non sur la théorie de l'interaction phénoménale que les phénoménologues ont en tête. Cependant, si l'on considère la conscience impliquant une expérience phénoménale, il semble difficile d'appliquer autre chose que cela

La théorie de l'esprit rend compte de l'empathie envers le HRI. En d'autres termes, le problème concernant la compatibilité des théories phénoménologiques avec l'HRI semble être l'aspect phénoménal des états mentaux, en particulier le côté ressenti et expérientiel des émotions. Alors que nous pourrions théoriser (TT) sur les composantes cognitives de, ou simuler, la situation de décision (ST) d'un robot, puis déduire ou projeter nos conclusions sur la situation du robot, il serait difficile de parler d'une compréhension empathique des capacités affectives du robot. et des états sensationnels de manière non projective. Si l'on étend le problème au concept d'« expérience » – terme central de l'approche phénoménologique (DPT) – les choses deviennent encore plus compliquées. Comme décrit ci-dessus, selon le DPT et ses variantes, dans notre interaction sociale avec les autres, nous percevons leurs expériences avec empathie, et nous le faisons à partir d'un point de vue réciproque à la deuxième personne. « L'expérience » est un terme phénoménologique d'élaboration et implique des aspects existentiels et des qualités phénoménologiques. Nous expérimentons subjectivement et consciemment notre monde ou ce que cela signifie de ressentir ou de faire quelque chose, par exemple percevoir une table rouge comme rouge et à quoi ressemble cette rougeur. La DPT suppose que nous vivons les expériences phénoménales des autres de manière directe et intersubjective, non pas en reproduisant le caractère qualitatif exact d'une expérience, mais plutôt en prêtant attention à la structure intentionnelle du point de vue de l'autre (Gallagher 2012 ; Zahavi et Michael 2018) . Pour que ce processus fonctionne, l'interaction intercorporelle et face-à-face est importante.16 Or, alors que cette dernière est (au moins très fondamentalement) garantie lorsque nous coopérons et collaborons avec des robots, il manque certains critères cruciaux de cette relation intersubjective. : tout comme les robots ne ressentent pas d'émotions, ils n'ont pas d'expérience subjective avec leur contenu phénoménal et leur impact existentiel ... ce que c'est pour eux.

Nous n'avons pas besoin de recourir à des inférences, des imitations ou des projections théoriques. Nous constatons que l'autre vit des expériences phénoménales. Cela dit, d'un point de vue phénoménologique, il semble difficile de sympathiser avec les robots. Pourtant, en comparant l'intelligence artificielle avec des personnages fictifs, je proposerai une solution potentielle et démontrerai également que non seulement nous lisons dans les pensées ou reflétons le comportement des robots, mais qu'il est possible, au moins dans une certaine mesure, d'appliquer une approche phénoménologique, c'est-à-dire, faire preuve d'empathie interactive avec « l'expérience » perspective des robots. Et l'argument va même au-delà de cette analogie : lorsque nous interagissons avec des robots dans un environnement partagé, nous développons une intentionnalité commune et même une histoire commune, ce qui est crucial pour notre relation avec les robots (Coeckelberg 2 Cependant, à l'instar de notre compréhension empathique des personnages de fiction, notre capacité d'imagination est ici cruciale.

Reprenons l'analogie : on suppose généralement que l'empathie joue un rôle essentiel dans nos relations avec des récits et des personnages fictifs - que ce soit dans un roman, un film ou une pièce de théâtre. Depuis les années 1990, il y a eu un débat considérable au sein de la philosophie de la littérature et du cinéma sur la question de savoir si « l'empathie » devait être englobée sous le terme générique d'« engagement émotionnel » avec les personnages de fiction en général (par exemple

[16] La version narrativiste des approches phénoménologiques implique cependant une composante imaginative qui nous permet de comprendre la structure intentionnelle par l'imagination narrative, par exemple si une interaction intersubjective n'est pas donnée (Gallagher et Gallagher 2019) .

[17] C'est une question similaire à celle de la soi-disant « pensée zombie », qui se demande si nous pouvons assumer ou attribuer une conscience dans le cas des zombies – qui sont comme nous à tous égards physiques mais n'ont pas d'expériences conscientes au sens riche du terme. (Chalmers 1996 ; Dennett 1991).

Plantationa 2009 ; Smith, 1995). D'autres formes d'engagement incluent la contagion émotionnelle et le partage émotionnel – en particulier en ce qui concerne les effets maussades d'une fiction – la sympathie ou la compassion, les émotions négatives telles que l'antipathie et l'affection synesthésique (Plantinga 2009 ; Schmetkamp 2017) . Comme l'ont souligné de nombreux spécialistes du cinéma, l'empathie joue un rôle épistémique crucial en permettant au spectateur de suivre le récit et de rester attaché aux personnages (Smith 1995), laissant de côté l' autre débat complexe concernant ce qu'on appelle le « paradoxe de la fiction » – qui discute de la question de savoir si nous pouvons ressentir des émotions réelles envers des entités fictives et si ces émotions sont rationnelles (Yanal 1999) – et en supposant que nous ressentons réellement et devons ressentir de l'empathie envers des personnages fictifs, nous devons encore expliquer la meilleure façon de conceptualiser l'empathie dans le cas de la fiction. . Même si je suis généralement convaincu que nous déployons différentes formes d'empathie, de lecture mentale et de compréhension (c'est-à-dire le spectre complet de la compréhension des états mentaux des autres) lorsque nous regardons un film ou lisons un roman, je suppose qu'un aspect est particulièrement vital. : Les personnages fictifs expriment et représentent certaines perspectives individuelles sur leur monde (fictif). Ces perspectives sont racontées dans le monde diégétique du film ou du roman ; De plus, ils sont souvent en outre formulés par un auteur implicite ou explicite. Ils sont intégrés dans un récit plausible. Ou, en d'autres termes : un récit est une représentation structurée et façonnée d'événements depuis une certaine perspective (Goldie 2012 : 8) et dans la fiction, les personnages incarnent, s'expriment et représentent de telles perspectives intégrées.

L'importance des perspectives pour la fiction, et en fait pour notre engagement empathique dans celle-ci, est en partie due au fait qu'une fiction a généralement (mais pas toujours) des perspectives techniques différentes : une histoire est généralement racontée à la première ou à la troisième personne. perspective. Mais plus important encore, une perspective est une vision du monde. Cela dit, une perspective signifie comment une personne est ancrée dans le monde, comment elle perçoit le monde, comment elle le vit. Cette perspective est façonnée par et, à son tour, façonne les émotions, les expériences, les histoires, les souvenirs ; il est influencé par et influence lui-même les traits de caractère, les jugements et les croyances (Schmetkamp 2017). Lorsque, par exemple, nous sommes d'humeur dépressive, nous voyons notre monde d'un point de vue différent – à savoir dépressif ou mélancolique – que si nous étions dans un état de bonheur.

On peut désormais parler de personnages de fiction comme « ayant » (ou plutôt exprimant et représentant) une perspective dans la mesure où ils sont focalisés et racontés par un narrateur qui construit et dirige leur vision du monde. En tant que lecteurs ou spectateurs, nous les regardons comme s'ils avaient une perspective et nous pouvons imaginer ce que cela pourrait être d'avoir une telle perspective. L'empathie envers les personnages fictifs implique une sorte de prise de perspective centrée sur autrui sans réduire ce processus à une simple simulation ou projection égocentrique.19 Qui plus est, c'est un avantage des récits fictifs qu'ils transmettent les perspectives des autres de manière condensée. Les fictions nous offrent l'opportunité de nous immerger dans des perspectives qui peuvent être similaires ou totalement différentes des nôtres, et elles le font souvent de manière intense, condensée et globale.

---

[18] L'empathie en tant que prise de perspective est en effet une capacité qui permet aux spectateurs de comprendre les récits et les perspectives des personnages. Cependant, en tant que forme de compréhension sensible des raisons pour lesquelles le personnage ressent, pense et agit comme il le fait, c'est aussi un résultat. Ainsi, Coplan (2011) et Goldie (2000) ont soutenu que l'empathie est à la fois un processus et un résultat .

[19] Misselhorn a avancé un argument similaire en notant qu'« en voyant le T-ing d'un objet inanimé, nous imaginons percevoir un T-ing humain » (2009 : 353).

En comparant les robots avec des personnages fictifs, une caractéristique centrale congruente ressort : les deux n'ont pas vraiment d'émotions ni de croyances conscientes, mais ils peuvent les exprimer et les représenter. Et c'est en partie sur cette base que nous, en tant que destinataires ou sympathisants, leur attribuons des états mentaux semblables à ceux des humains (Weber 2013). Cependant, nous les vivons également comme des entités incarnées avec lesquelles nous interagissons. Comme l'a soutenu la philosophe phénoménologique du cinéma Vivian Sobchack, le film et ses personnages ne sont pas de simples projets ; ils ont un corps et une voix, et permettent des expériences quasi intersubjectives entre eux et les destinataires (Sobchack 2004). Ils pourraient même permettre des impressions tactiles. Cette caractéristique incarnée est également vraie pour les robots, peut-être même plus encore.

Il existe néanmoins des différences cruciales. Premièrement, contrairement aux robots, les personnages de fiction manquent d'une capacité vitale pour toute explication intersubjective de l'empathie : à savoir l'interaction réciproque. Dans nos relations avec des personnages fictifs, nous devons imaginer que les personnages expriment les émotions, les expériences et les perspectives, mais nous n'interagissons pas réciproquement avec eux. De plus, les personnages fictifs ne peuvent pas opposer leur veto à tout ce que nous leur attribuons. En revanche, dans nos rencontres avec les robots, il existe au moins une entité existante et présente, incarnée et intégrée, en interaction avec laquelle nous pouvons développer une relation. Le robot est capable de s'opposer à quelque chose - par exemple, si j'étais un patient et que je ne voulais pas prendre mes médicaments, le robot pourrait être chargé de veiller à ce que je le fasse. Deuxièmement, on pourrait objecter que, contrairement aux personnages de fiction, les robots n'ont pas (encore) de perspective expérientielle ni de récit individuel, comme évoqué plus haut. Les fictions offrent en effet une riche image de la manière dont une personne peut percevoir et évaluer son monde ; Et grâce à ces cadres et pratiques narratifs, nous élargissons notre horizon et apprenons de nouvelles émotions ou nuances émotionnelles. Cependant, les émotions et les expériences des personnages fictifs ne sont également racontées que dans un cadre narratif particulier ; Leur développement dépend à la fois de ce qu'un narrateur a conçu dramaturgiquement et de la façon dont les lecteurs ou les spectateurs le reçoivent par rapport à leur propre bagage intellectuel et expérientiel. Les émotions et expériences fictives ont moins de flexibilité et de créativité que leurs homologues humaines. Cela dit, la question se pose de savoir si les personnages de fiction peuvent encore être comparés aux robots. Les personnages fictifs ne vivent vraiment rien ; De même, les robots n'ont pas d'expériences riches et qualitatives. Cependant, les robots perçoivent au moins leur environnement, le catégorisent, l'évaluent et interagissent au sein de celui-ci. Ils ont une façon de voir et d'être au monde ; ils sont incarnés et contextualisés. Si l'on pense au célèbre exemple anti-réductionniste de Thomas Nagel « Qu'est-ce que ça fait d'être une chauve-souris ? (Nagel 1974) Nous ne serons jamais capables de comprendre entièrement la perspective expérientielle des autres êtres ; une chauve-souris, du moins c'est ce que dit son argument, a un système de perception totalement différent qui ne peut être comparé à la perception humaine. Pourtant, les scientifiques découvrent en permanence de nouveaux faits sur les entités non humaines comme les poissons ou les plantes (Coeckelberg 2018 : 148), et un argument ici soutient que même si nous ne saurons peut-être jamais ce que c'est que d'être avec eux, nous pouvons au moins expérimenter eux et leur point de vue dans notre relation avec eux (i

Si nous essayons de comparer le point de vue du robot avec le nôtre, nous constatons certaines similitudes, mais aussi bien sûr de nombreuses différences. Mais il ne s'agit pas d'un phénomène nouveau dans notre cognition sociale des autres esprits. Premièrement, un robot perçoit le monde d'une certaine manière littéralement (par exemple visuellement) (peut-être de manière humaine, peut-être pas). Deuxièmement, en tant qu'intelligence artificielle, elle a également une perspective dans le sens où elle perçoit et évalue le monde qui l'entoure, la manière dont elle résout les problèmes, etc. La perspective du robot est loin d'être une perspective au sens élaboré, comme celle des êtres humains, mais c'est une perspective épistémique et évaluative : un robot sait quelque chose et porte des jugements sur le monde. On peut aussi affirmer qu'il a une perspective motivationnelle, car un robot agit sur lui

base de ses croyances.20 Plus important encore, les robots sont intégrés dans un contexte que nous percevons ou avec lequel nous interagissons. Donc, ma réponse à la question de savoir si nous pouvons sympathiser avec les robots est : oui. De plus, tous les comptes existants sont plus ou moins applicables à HRI. Bien sûr, la prochaine question que nous devons nous poser est la suivante : devrions-nous le faire ?

3 Devrions-nous sympathiser avec les robots ?

Compte tenu de l'analyse précédente, supposons que nous pouvons sympathiser avec les robots humanoïdes de plusieurs manières, c'est-à-dire que nous pouvons ressentir, interagir avec ou déduire de leurs « croyances », « émotions », « expériences » et « perspectives ». Mais pourquoi devrions-nous sympathiser avec eux ? À la lumière de l'utilisation croissante des robots dans les domaines de la médecine, de la santé et des soins aux personnes âgées, par exemple, il semble beaucoup plus plausible que les robots sympathisent avec les patients plutôt que l'inverse. Ils doivent, d'une manière ou d'une autre, engager certaines sensibilités à l'égard des besoins des patients, tandis que, à leur tour, les patients humains pourraient avoir besoin d'un compagnon empathique. Cela dit, il semble que l'enquête menée jusqu'à présent ait été avant tout un test théorique visant à révéler lesquels des différents récits d'empathie sont compatibles avec l'HRI. Mais y a-t-il aussi une raison pour laquelle nous, en tant qu'humains, devrions également sympathiser avec les robots ? Cette question est pertinente, car les relations entre humains et robots ne sont réussies et fructueuses que si les deux interagissent effectivement l'un avec l'autre, et ces interactions peuvent présupposer – d'une manière ou d'une autre – un engagement empathique.

Trois arguments pourraient être avancés en faveur de cette thèse normative :

1. Un argument stratégique ; 2.Un argument anti-barbarisation ; 3. Un argument pragmatique et communautaire partagé.

Le premier argument, stratégique, n'est pas directement un argument normatif moralement pertinent. Il reprend l'idée que pour interagir avec succès, nous devons d'une manière ou d'une autre être capables de déduire et de comprendre ce que fait notre homologue interactif. Plus précisément, nous pourrions vouloir faire preuve d'empathie, adopter une perspective ou lire dans une autre pensée afin de mieux atteindre nos objectifs. Notre interaction avec les robots et notre empathie à leur égard ne servent en ce sens qu'à autre chose ; c'est simplement instrumental. La notion « devrait » fait référence à un impératif hypothétique. À cet égard, les robots sont davantage considérés comme des outils que comme des collaborateurs. En fait, ils ne sont pas ici considérés comme des agents moraux ou des patients, qui ont un statut moral (Coeckelberg 2018).

Plus substantiel et moralement normatif est le deuxième argument de non-barbarisation ou de culture. En ne faisant pas preuve d'empathie envers les autres, dit-on, nous risquons de devenir désensibilisés. À son tour, l'empathie pourrait cultiver un comportement prosocial et améliorer notre caractère moral. Avant d'explorer les principaux problèmes de cette thèse, j'expliquerai deux de ses racines – à savoir un argument kantien et un argument aristotélicien. L'argument kantien a été avancé à l'origine à propos de la relation homme-animal. Ça implique

[20] Là encore, des arguments similaires pourraient être avancés pour d'autres formes d'IA d'agents non humains, par exemple des formes virtuelles abstraites. Cet article se concentre sur les robots humanoïdes avec lesquels les êtres humains coopèrent et collaborent. Pour que cela réussisse, les êtres humains pourraient attribuer à l'IA non seulement des états mentaux de base, mais aussi une perspective et un récit. Cela pourrait être important pour l'intentionnalité et l'attention collectives.

que nous ne devrions pas être cruels envers les animaux car cela nuirait ou corromprait notre caractère moral en général. Selon cet argument, les animaux ne sont que des patients moraux indirects, sans avoir de statut moral propre, puisque Kant lie le statut moral d'une personne à la compétence d'agir de manière autonome à partir de la raison et attribue cette compétence uniquement aux personnes. Le même argument serait alors valable pour les robots sociaux qui ne seraient pas en soi des destinataires moraux : c'est parce qu'ils pourraient ne pas avoir d'autonomie au sens élaboré. Cependant, en ne sympathisant pas avec eux, nous manquerions de respect à une condition cruciale de l'humanité … notre moralité – si nous traitons les animaux de manière inhumaine, nous devenons des personnes inhumaines. Cela s'étend logiquement au traitement des compagnons robotiques. […] Cela peut également empêcher la désensibilisation à l'égard des créatures vivantes réelles et protéger l'empathie que nous avons les uns envers les autres » (Darling 2016 : 19).

L'argumentation aristotélicienne va dans le même sens. C'est que nous pouvons cultiver nos émotions en prenant du recul, définissant ainsi la perspective comme une prise de distance par rapport à notre propre point de vue à la première personne, ou en partageant des émotions et en nous familiarisant ainsi avec de nouvelles émotions (Nussbaum 2011 ; Rorty 2001) . Alors que la vision kantienne insiste sur le problème de la barbarisation, la vision aristotélicienne souligne l'impact éthique du fait de cultiver quelque chose en faisant preuve d'empathie : nos émotions, notre perception morale, notre imagination et notre pouvoir de jugement.

Comme je l'ai dit, certains problèmes se posent ici, qui assaillent particulièrement la vision kantienne : le premier est que la reconnaissance d'un simple statut indirect de non-personnes ou d'êtres sans « rationalité » n'est pas satisfaisante : c'est contre-intuitif, anthropocentrique, et cela exclut bien plus que les entités non humaines (Gruen 2017). Mais cela concerne-t-il également les entités inanimées ? Ainsi, la question demeure : à quoi faisons-nous du mal lorsque nous utilisons la violence contre des robots qui pourraient ne rien ressentir de manière élaborée et subjective ? Ont-ils une conception du respect et de la dignité ? Ont-ils des revendications morales ? Ces questions complexes resteront sans réponse ici, car elles nécessiteraient leur propre enquête. Une autre objection contre le point de vue kantien est que l'argument repose sur une vision spécifique de l'empathie en tant que comportement prosocial. Cela implique non seulement une compréhension des autres esprits, mais également le souci du bien-être d'une autre entité. Autrement dit, l'empathique ne s'intéresse pas seulement aux expériences de l'autre et « les ressent » ; ils sont également motivés à soulager la souffrance de l'autre ou à favoriser son bien-être. Et si nous étions cruels envers eux et manquions de respect pour leur bien-être – par exemple en les battant ou en les violant (si nous pensons aux robots sexuels) – cela se répercuterait également sur notre comportement envers les humains. Cependant, comme indiqué précédemment, l'impact éthique de l'inquiétude ou de l'attention est plutôt l'impact de la sympathie ou de la compassion en tant qu'émotion morale sui generis, et en tant que telle, elle est distincte de l'empathie (Darwall 1998) . Comme l'ont notamment montré les phénoménologues, l'empathie n'est pas nécessairement une attitude positive envers autrui, mais peut aussi conduire à un comportement antisocial.
Une personne sadique doit également être empathique dans ce sens, c'est-à-dire qu'elle comprend la souffrance de l'autre mais ne veut pas la soulager (Breithaupt 2019 ; Zahavi et Michael

[21] Kant écrit : « Si un homme tire sur son chien parce que l'animal n'est plus capable de rendre service, il ne manque pas à son devoir envers le chien, car le chien ne peut pas juger, mais son acte est inhumain et porte atteinte en lui à cette humanité qu'il est son devoir de le montrer envers l'humanité. S'il ne veut pas étouffer ses sentiments humains, il doit faire preuve de bonté envers les animaux, car celui qui est cruel envers les animaux devient dur aussi dans ses relations avec les hommes » (Kant 1997 : 212) .

2018).22 En d'autres termes, une approche kantienne amalgame certaines différenciations conceptuelles importantes, notamment entre l'empathie et la compassion. Une autre objection pourrait être soulevée ici : empiriquement, il n'est pas du tout clair pourquoi quelqu'un qui ne sympathise pas avec les autres devient nécessairement barbarisé (Brinck et Balkenius 2018).

Cependant, d'un point de vue plus optimiste, certains soutiennent qu'une compréhension empathique fréquente ou une prise de perspective peut nous aider à apprendre ce que les autres pourraient ressentir ou penser.
Plus nous déployons de l'empathie, plus nous sommes capables de nous impliquer avec les autres aussi bien dans nos interactions quotidiennes que dans des rencontres plus insolites. De plus, cela pourrait faire de nous une personne plus tolérante ou plus vertueuse. Encore une fois, cela est argumenté en ce qui concerne les personnages et les récits fictifs. Prêter attention aux points de vue et aux expériences des autres est, comme l'a affirmé Richard Rorty, d'une valeur éthique, puisque ce faisant, nous abandonnons notre perspective égocentrique (Rorty 2001). Mais, bien sûr, nous pourrions adopter cet argument en faveur de l'HRI : faire preuve d'empathie envers les robots améliorerait alors nos interactions coopératives et collaboratives dans la mesure où nous les connaîtrions mieux. Cela conduit au troisième argument qui partage certaines caractéristiques à la fois avec l'approche stratégique et avec l'approche kantienne/aristotélicienne, mais met l'accent sur l'interaction, la relation et la compréhension sociale de soi des empathisants.

Cet argument (qui décrit ma propre position) reprend l'approche de Rorty mais la modifie pour en faire une thèse encore plus pragmatiste et relationnelle de la cognition sociale et de ses conditions préalables. Contrairement à l'approche kantienne et aristotélicienne, cette vision part d'un point de vue anti-anthropocentrique et met l'accent sur la relation interactive entre les êtres humains et les robots. Cette position suppose que l'empathie envers les autres - dans toutes ses variantes mais surtout dans la tradition phénoménologique interactive - peut nous permettre de nous familiariser avec « l'être au monde » des autres et ainsi d'élargir nos horizons, de changer nos perspectives et de façonner nos interactions sociales. … et le comportement moral envers les autres non-humains.

Compte tenu des perspectives, je suppose ici que nous pouvons même parler de (futurs) robots et systèmes d'apprentissage profond23 comme ayant une vision spécifique du monde qui leur est propre.
Cette vision sera à certains égards similaire et à d'autres égards différente des perspectives humaines. Des films de science-fiction comme HER (États-Unis, 2013) ont imaginé ce que pourrait devenir l'IA indépendante : des systèmes superintelligents qui dépassent de loin les capacités de la pensée humaine. Faire preuve d'empathie envers les robots humanoïdes avec lesquels nous interagissons de plus en plus - dans le contexte des soins de santé, par exemple - pourrait nous aider à nous préparer aux développements futurs. Pour l'instant, cependant, dans la mesure où nous partageons déjà des actions et des environnements avec des robots, et dans la mesure où l'empathie et la cognition sociale peuvent améliorer nos interactions avec les autres, nous pouvons également supposer que nos interactions avec les robots bénéficieront de un point de vue empathique, mais pas simplement dans un sens instrumental et stratégique. Cela pourrait également avoir un effet formateur, un argument qui, comme nous l'avons noté, a également été avancé à propos des mondes fictifs. Mais le point le plus important est qu'une telle vision touche à la question de savoir comment nous voulons nous comprendre : prendre au sérieux les robots en tant que compagnons sociaux devrait être mis en œuvre dans le cadre de notre compréhension de nous-mêmes en tant qu'humains et membres de sociétés démocratiques.

La façon dont nous interagissons avec les robots dépend beaucoup de la façon dont nous les considérons : comme avec les outils

22 Le phénomène selon lequel les sympathisants peuvent devenir encore plus cruels à mesure que les robots ressemblent davantage à des humains sont appelés les « vallée étrange » (voir Misselhorn 2009 ; Mori 2005).

23 Ou comme les appelle Susan Schneider : « les esprits du futur » (sous presse).

qui sont censés interagir d'un simple point de vue instrumental, ou en tant que partenaires que nous devrions prendre au sérieux pour leur propre bien. C'est donc la relation et la communauté partagée qui sont ici mises en avant. Une telle position souligne l'impact pragmatiste et phénoménologique des interactions. Cela pourrait également avoir des implications sur le statut des robots en tant qu'agents moraux et patients moraux, comme le soutient Mark Coeckelbergh : « La question du statut moral est toujours liée à la question de savoir qui fait partie de la communauté morale et à quels jeux moraux sont déjà joués. » (Coeckelberg 2018 : 149).

Au lieu d'une mise en œuvre descendante de la moralité, Coeckelbergh plaide en faveur d'une perspective ascendante. En considérant les robots comme des compagnons dans un contexte relationnel et en faisant preuve d'empathie avec leur récit de perspective, nous développons une relation avec eux qui, à son tour, a des effets sur la façon dont nous les percevons moralement (ibid.).24 Cependant, discuter du statut moral irait au- delà la portée de cet article. Comme mentionné ci-dessus, l'empathie n'est pas en soi une émotion morale ou une attitude de bienveillance. Mais elle pourrait semer les graines pertinentes à cet égard, puisqu'elle fournit la base épistémologique d'une morale intersubj
De plus, cela a beaucoup à voir avec notre compréhension sociale et morale de nous-mêmes : « [La] façon dont nous traitons les autres entités, la façon dont nous les vivons, ce que nous disons à leur sujet, la façon dont nous les traitons, etc. en dit aussi beaucoup sur moi et en dit long sur nous. (Coeckelberg 2018 : 150). Mais au lieu d'une vision anthropocentrique, il s'agit plutôt d'une vision relationnelle qui traite les entités non humaines comme des partenaires en interaction.

4. Conclusion

L'intelligence artificielle en général et les robots humanoïdes en particulier vont changer nos vies et peut-être nous-mêmes aussi. Les philosophes ont beaucoup à considérer en termes d'impacts épistémiques, éthiques, esthétiques et politiques de ces nouveaux défis. L'empathie n'est que l'un des nombreux sujets remis en question par HRI. Cet article a contribué aux investigations nécessaires déjà en cours ou à venir. J'ai discuté du casse-tête épistémique de savoir si nous pouvons sympathiser avec les robots, en appliquant les récits contemporains dominants de l'empathie à ce domaine. J'ai ensuite examiné la question normative de savoir si et pourquoi nous devrions sympathiser avec les robots. L'article propose un point de vue pragmatique en démontrant que a) nous pouvons effectivement sympathiser avec les robots humanoïdes, non seulement à un niveau basique, mais aussi, au moins dans une certaine mesure, à un niveau de prise de perspective imaginative ; De plus, il a été démontré que même d'un point de vue phénoménologique et intersubjectif, il est possible de parler d'empathie avec les robots intégrés dans notre monde, avec lesquels nous interagissons et partageons un récit contextuel. L'accent est mis sur l'empathie en tant que processus d'interaction mutuelle plutôt que comme résultat. Cependant, l'article soutient également que b) nous devrions sympathiser avec les robots humanoïdes, car ce faisant, nous pouvons acquérir de nouvelles connaissances sur un être-dans-le-monde très inconnu, élargissant ainsi nos horizons, nous formant aux futurs développements de l'IA et améliorant l'HRI dans un environnement social partagé. Cela a été jugé non seulement d'une valeur instrumentale, mais également précieux pour notre compréhension de nous-mêmes et de notre société dans laquelle les robots et autres formes d'IA peuvent être considérés comme des compagnons.

[24] Coeckelbergh propose une approche similaire à la mienne mais s'inspire des concepts de forme de vie et de jeux de langage de Wittgenstein. Pourtant, son article manque d'une définition claire de ce qu'il pense que l'empathie implique (par exemple si l'empathie implique effectivement de se soucier du bien-être de l'autre, comme son article semble le suggérer).

Les références

Baron-Cohen, S. 1995. Cécité mentale. Un essai sur l'autisme et la théorie de l'esprit. Cambridge, MA : MIT Press.

Batson, CD 2009. Ces choses appelées empathie : huit phénomènes liés mais distincts. Dans Les neurosciences sociales de l'empathie, éd. J. Decety et W. Ickes, 3-15. Cambridge, MA : MIT Press.

Benford, G. et E. Malartre. 2007. Au-delà de l'humain. Tom Doherty Associates : Vivre avec des robots et des cyborgs. New York.

Boddington, P., P. Millican et M. Wooldridge. 2017. Numéro spécial Esprits et machines : Éthique et intelligence artificielle. Esprits et machines 27(4) : 569-574.

Boden, MA 2016. IA. Sa nature et son avenir. Oxford : Presse universitaire d'Oxford.

Breazeal, CL 2002. Conception de robots sociaux. Cambridge, MA : MIT Press.

Breithaupt, F. 2019. Les côtés obscurs de l'empathie. Ithaque : Cornell University Press.

Bretan, M., G. Hoffman et G. Weinberg. 2015. Comportements physiques dynamiques et émotionnellement expressifs chez les robots. Revue internationale d'études humain-informatique 78 : 1–16.

Brinck, I. et C. Balkenius. 2018. Reconnaissance mutuelle dans l'interaction homme-robot : un récit déflationniste. Philosophie et technologie : 1–18. https://doi.org/10.1007/s13347-018-0339-x.

Chalmers, DJ 1996. L'esprit conscient. Oxford : Presse universitaire d'Oxford.

Coeckelberg, M. 2018. Pourquoi se soucier des robots ? Empathie, position morale et langage de la souffrance. Caïros. Journal de philosophie et de science 20 : 141-158.

Colombetti, G. 2013. Le corps sensible. La science affective rencontre l'esprit actif. Cambridge, MA : MIT Press.

Coplan, A. 2011. Comprendre l'empathie, 3-18. Ses caractéristiques et ses effets. En empathie. Philosophique et perspectives psychologiques. Oxford : Presse universitaire d'Oxford.

Coplan, A. et P. Goldie. 2011. Empathie. Perspectives philosophiques et psychologiques. Oxford : Oxford Presse universitaire.

Cross, ES, Riddoch, KA, Pratts, J, Titone, S, Chaudhury, B et Hortensius, R. 2018. Une enquête neurocognitive sur l'impact de la socialisation avec un robot sur l'empathie pour la douleur. Préimpression. https://doi. org/10.1101/470534.

Darling, K. 2016. Étendre la protection juridique aux robots sociaux : les effets de l'anthropomorphisme, de l'empathie et des comportements violents envers les objets robotiques. Dans Droit des robots, éd. M. Froomkin, R. Calo et I. Kerr. Cheltenham : Edward Elgar.

Darwall, S. 1998. Empathie, sympathie, attention. Études philosophiques 89 : 261-282.

De Sousa, R. 1987. La rationalité de l'émotion. Cambridge, MA : MIT Press.

De Vignemont, F. et P. Jacob. 2012. Qu'est-ce que ça fait de ressentir la douleur d'autrui ? Philosophie des sciences 79 (2) : 295-316.

De Vignemont, F. et T. Singer. 2006. Le cerveau empathique : Comment, quand et pourquoi ? Tendances cognitives sciences 10(10) : 435-441.

Dennett, D. 1991. La conscience expliquée. Boston : Little, Brown et Co.

Dullstein, M. 2012. La deuxième personne dans le débat sur la théorie de l'esprit. Revue de Philosophie et Psychologie 3 (2) : 231-248.

Dullstein, M. 2013. Perception directe et simulation : le récit de Stein sur l'empathie. Revue de philosophie et de psychologie 4 : 333-350.

Dumouchel, P. et L. Damiano. 2017. Vivre avec des robots. Cambridge, MA : Presse universitaire de Harvard.

Engelen, EM 2018. Pouvons-nous partager un sentiment de nous-mêmes avec une machine numérique ? Le partage émotionnel et le reconnaissance de l'un comme de l'autre. Revues scientifiques interdisciplinaires 43(2) : 125-135.

Engelen, E.M. et B. Röttger-Rössler. 2012. Débats disciplinaires et interdisciplinaires actuels sur l'empathie. Revue des émotions 4 (1) : 3–8.

Fodor, J. 1987. Psychosémantique. Le problème du sens dans la philosophie de l'esprit. Cambridge, Massachusetts : MIT Presse.

Gallagher, S. 2008. Perception directe dans le contexte interactif. Conscience et cognition 17 (2) : 535–543.

Gallagher, S. 2017. Empathie et théories de la perception directe. Dans Le manuel Routledge de philosophie de empathie, éd. H. Maibom, 158-168. New York : Routledge.

Gallagher, S. et J. Gallagher. 2019. Agir comme un autre : l'empathie d'un acteur pour son personnage. Topoï (en ligne d'abord), https://doi.org/https://doi.org/10.1007/s11245-018-96247.

Gallagher, S. et D. Hutto. 2008. Comprendre les autres à travers l'interaction primaire et la pratique narrative. Dans L'esprit partagé : perspectives sur l'intersubjectivité, éd. J. Zlatev, T. Racine, C. Sinha et E. Itkonen, 17-38. Amsterdam/Philadelphie : John Benjamins Publishing Company.

Gallese, V. 2001. L'hypothèse du « collecteur partagé » : des neurones miroirs à l'empathie. Journal de
Études de conscience 8 : 33-50.

Goldie, P. 2000. Les émotions. Oxford : Presse universitaire d'Oxford.

Goldie, P. 2012. Le désordre à l'intérieur. Récit, émotion et esprit. Oxford : Presse universitaire d'Oxford.

Goldman, A. 2006. Simulation des esprits : philosophie, psychologie et neurosciences de la lecture mentale. Oxford :
Presse de l'Université d'Oxford.

Goldman, A. 2011. Deux voies vers l'empathie : aperçus des neurosciences cognitives. Dans Empathie : perspectives philosophiques et
psychologiques, éd. A. Coplan et P. Goldie, 31-44. Oxford : Presse universitaire d'Oxford.

Gopnik, A. et H.M. Wellman. 1994. La théorie théorique. Dans Cartographier l'esprit : spécificité du domaine dans la cognition et la culture,
éd. LA Hirschfeld et SA Gelman, 257-293. Cambridge : La Presse de l'Universite de Cambridge.

Gruen, L. 2009. S'occuper de la nature : engagement empathique avec le monde plus qu'humain. L'éthique et le
Environnement 14 (2) : 23-38.

Gruen, L. 2017. Le statut moral des animaux. Dans L'Encyclopédie de philosophie de Stanford (édition automne 2017),
éd. FR Zalta, https://plato.stanford.edu/archives/fall2017/entries/moral-animal/.

Hickok, G. 2014. Le mythe des neurones miroirs : la véritable neuroscience de la communication et de la cognition. Nouveau
York : WW Norton & Company.

Hoffmann, M. et R. Pfeifer. 2018. Les robots comme alliés puissants pour l'étude de la cognition incarnée de bas en haut. Dans Le manuel
d'Oxford de la cognition 4E, éd. A. Newen, L. de Bruin et S. Gallagher.
Oxford : Presse universitaire d'Oxford.

Hutto, DD 2008. L'hypothèse de la pratique narrative : clarifications et implications. Explorations philosophiques 11(3) : 175-192.

Iacoboni, M. 2011. Les uns dans les autres : mécanismes neuronaux de l'empathie dans le cerveau des primates. Dans Empathie :
perspectives philosophiques et psychologiques, éd. A. Coplan et P. Goldie, 45-57. Oxford : Presse universitaire d'Oxford.

Iacoboni, M., RP Woods et al. 1999. Mécanismes corticaux de l'imitation humaine. Sciences 286 : 2526-2528.

Kanske, P. 2018. L'esprit social : Démêler les méthodes affectives et cognitives pour comprendre les autres.
Revues scientifiques interdisciplinaires 43 (2) : 115-124.

Kant, I. 1997. Conférences sur l'éthique, éd. et trad. P. Heath et JB Schneewind. Cambridge : Cambridge
Presse universitaire.

Kasparov, G. 2017. Réflexion profonde : là où s'arrête l'intelligence artificielle et où commence la créativité humaine. New York:
Affaires publiques.

Leite, A., A. Pereira, S. Mascarenhas, C. Martinho, R. Prada et A. Paiva. 2013. L'influence de l'empathie dans les relations homme-robot.
Revue internationale d'études humain-informatique. 71 (3) : 250-260.

Lin, P., R. Jenkins et K. Abney. 2017. Éthique du robot 2.0 : De la voiture autonome à l'intelligence artificielle.
Oxford : Presse universitaire d'Oxford.

Loh, J. 2019. Roboterethik. Voici ce dont vous avez besoin. Berlin : Suhrkamp.

MacLennan, B.J. 2014. Traitement éthique des robots et problème difficile des émotions des robots. Journal international des émotions
synthétiques 5 (1) : 9-16.

Maibom, H. 2017. Le manuel Routledge de philosophie de l'empathie. Londres : Routledge.

Misselhorn, C. 2009. Empathie avec les objets inanimés et l'étrange vallée. Esprits et machines 19 (3) :
345-359.

Misselhorn, C. Sous presse. L'empathie envers les robots est-elle moralement pertinente ? Dans Machines émotionnelles : perspectives de
l'informatique affective et de l'interaction émotionnelle homme-machine, éd. C. Misselhorn et M. Klein.
Wiesbaden.

Mori, M. 2005. Sur l'étrange vallée. Dans Actes de l'atelier Humanoïdes-2005 : Vues de l'étrange vallée. Tsukuba : Japon.

Nagel, T. 1974. Qu'est-ce que ça fait d'être une chauve-souris ? La revue philosophique 83 (4) : 435-450.

Newen, A. 2015. Comprendre les autres : la théorie du modèle de personne. Dans Open MIND : 26(T), éd. T. Metzinger
et JM Windt. Francfort-sur-le-Main : Groupe MIND.

Newen, A., L. De Bruin et S. Gallagher. 2018. Le manuel d'Oxford sur la cognition 4E. Oxford : Oxford
Presse universitaire.

Nussbaum, M. 2011. Bouleversements de la pensée : L'intelligence des émotions. Cambridge : Université de Cambridge
Presse.

Plantinga, C. 2009. Des spectateurs en mouvement : le cinéma américain et l'expérience du spectateur . Berkeley : Université de
Presse californienne.

Rorty, R. 2001. Rédemption de l'égoïsme : James et Proust comme exercices spirituels. Télos 3 (3) : 243-263.

Scheutz, M. 2011. Rôles architecturaux de l'affect et comment les évaluer chez les agents artificiels. Journal international des émotions
synthétiques 2 (2) : 48-65.

Schmetkamp, S. 2017. Gagner des perspectives sur nos vies : humeurs et expérience esthétique. Philosophie 45(4) : 1681-1695.

Schmetkamp, S. 2019. Theorien der Empathie - Ein Einführung. Hambourg : Junius Publisher.

Schneider, S. Sous presse. Esprits du futur : améliorer et transcender le cerveau.

Slote, M. 2017. Les nombreux visages de l'empathie. Philosophie 45 (3) : 843-855.

Smith, M. 1995. Personnages attachants : fiction, émotion et cinéma. Oxford : Presse Clarendon.

Sobchack, V. 2004. Pensées charnelles : incarnation et culture des images en mouvement. Berkeley : Université de Presse californienne.

Stein, E. 1989. Sur le problème de l'empathie : Les œuvres rassemblées d'Edith Stein. Vol. 3 (3e édition révisée), trans. W. Stein. Washington, DC : Publications ICS.

Stueber, K. 2006. Redécouvrir l'empathie : agence, psychologie populaire et sciences humaines. Cambridge, Massachusetts : Presse du MIT.

Stueber, K. 2018. Empathie. Dans L'Encyclopédie de philosophie de Stanford (édition printemps 2018), éd. FR Zalta, https://plato.stanford.edu/archives/spr2018/entries/empathy/.

Vaage, MB 2010. Film de fiction et variétés d'engagement empathique. Études du Midwest en philosophie 34 : 158-179.

Vallor, S. 2011. Carebots et soignants : maintenir l'idéal éthique des soins au 21e siècle. Philosophie et Technologie 24 (3) : 251-268.

Weber, K. 2013. Qu'est-ce que ça fait de rencontrer un agent artificiel autonome ? IA et SOCIÉTÉ 28 : 483-489.

Yanal, RJ 1999. Paradoxes de l'émotion et de la fiction. Pennsylvanie : Penn State University Press.

Zahavi, D. 2011. Empathie et perception sociale directe : une proposition phénoménologique. Revue de philosophie et de psychologie 2 (3) : 541-558.

Zahavi, D. 2014. Soi et les autres : explorer la subjectivité, l'empathie et la honte. Oxford : Université d'Oxford Presse.

Zahavi, D. et J. Michael. 2018. Au-delà du miroir : perspectives 4E sur l'empathie. Dans Le manuel d'Oxford de la cognition 4E, éd. A. Newen, L. de Bruin et S. Gallagher, 589-606. Oxford : Presse universitaire d'Oxford.