Dialogue (2025), **64**, 65–79
doi:10.1017/S0012217325000174

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Numéro spécial sur la normativité. À la recherche des normes perdues

Peut-il y avoir agentivité morale sans sentience ?

Thomas Emmaüs Adetou

Département de philosophie et Centre de recherche en éthique (CRÉ), Université de Montréal, Montréal, QC, Canada
Courriel : amah.adetou@umontreal.ca

Résumé

Carissa Véliz (2021) soutient que pour être un agent moral, une entité (comme un système algorithmique) doit nécessairement être sentiente. En prenant l'exemple d'une forme d'agentivité sans conscience phénoménale et donc sans sentience, je conteste cette thèse. En effet, certains groupes humains peuvent être considérés comme des agents intentionnels non réductibles aux individus particuliers qui les composent. En m'appuyant sur l'interprétativisme proposé par Uriah Kriegel (2011) et John Haugeland (1990), ainsi que sur la posture intentionnelle (intentional stance) de Daniel Dennett (1987), je montre que l'intentionnalité est une condition suffisante pour parler d'agentivité morale et que, contrairement à ce qu'affirme Véliz, il peut donc y avoir des agents moraux qui ne sont pas sentients.

Abstract

Carissa Véliz (2021) argues that for an entity (such as an algorithmic system) to be a moral agent, it must necessarily be sentient. By considering a form of agency that lacks phenomenal consciousness and, therefore, sentience, I challenge this thesis. Indeed, certain human groups can be regarded as intentional agents that are not reducible to the individual members composing them. Drawing on the interpretivist approaches proposed by Uriah Kriegel (2011) and John Haugeland (1990), as well as Daniel Dennett's (1987) intentional stance, I argue that intentionality constitutes a sufficient condition for moral agency. Consequently, contrary to Véliz's claim, moral agents need not necessarily be sentient.

Mots-clés : sentience ; conscience ; agentivité ; agentivité intentionnelle ; agentivité morale

Carissa Véliz (2021) défend l'idée qu'une entité, telle qu'un système algorithmique, ne peut être considérée comme un agent moral sans être sentiente. Le présent article remet en question cette affirmation en explorant une forme d'agentivité dénuée de

© The Author(s), 2025. Published by Cambridge University Press on behalf of the Canadian Philosophical Association/Publié par Cambridge University Press au nom de l'Association canadienne de philosophie. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

conscience phénoménale. L'exemple de certains groupes humains, perçus comme des agents intentionnels distincts des individus qui les composent, servira d'appui à cette critique. L'article se structure comme suit : je commence par présenter l'argument de Véliz dans la première section. Dans la deuxième partie, je propose une objection à l'une des prémisses de son argument. J'adresse ensuite une objection possible à ma position dans la troisième section. J'évalue enfin l'ensemble de ma démarche en guise de conclusion.

1. L'argument de la sentience de Véliz (2021)

L'argument de Carissa Véliz se situe dans les débats en éthique de l'intelligence artificielle (IA). Il a pour objectif, selon les affirmations de l'autrice, de s'opposer à un courant de pensée qui soutient que les systèmes d'IA ont un certain degré d'agentivité morale. Dans cette première section, je présente l'argument de Véliz, je le situe ensuite sommairement dans la littérature en éthique de l'intelligence artificielle, avant d'en présenter les prémisses et la conclusion.

L'argument de Véliz pourrait se reconstruire comme suit :

P [1] La sentience est une condition nécessaire pour être un agent moral.

P [2] Les zombies moraux et les algorithmes ne sont pas sentients.

C [3] Donc, les zombies moraux et les algorithmes ne sont pas des agents moraux.

En éthique de l'IA et plus précisément en éthique des machines, plusieurs auteurs réfléchissent à la possibilité d'une agentivité morale des agents artificiels (Müller, 2021 ; Anderson et Anderson, 2011 ; Wallach et Allen, 2009 ; Formosa et Ryan, 2021 ; Sharkey, 2020).

C'est dans cette optique que James Moor distingue quatre degrés d'agentivité morale des robots. Le premier, composé d'*agents ayant un impact moral*, regroupe des agents moraux au sens le plus minimal. Ce sont des agents dont les actions, qu'elles soient intentionnelles ou non, ont des conséquences sur le plan moral. Ainsi, l'utilisation des robots jockeys (« camels jockeys ») dans les courses de chameaux au Qatar a pour conséquence de réduire l'exploitation des enfants auparavant utilisés pour monter les chameaux. Au deuxième degré, Moor place les *agents moraux implicites*, c'est-à-dire les agents construits avec des règles morales implicites, notamment des considérations de sécurité ou de sûreté (par exemple, un système de sûreté intégré à la conduite autonome). Au troisième degré, nous avons des *agents moraux explicites*. Ce sont des agents qui agissent à partir de règles morales et qui peuvent prendre des décisions dans des situations morales diverses et complexes à partir de procédures éthiques raisonnables du point de vue normatif. En dernier lieu, il y a les *agents pleinement moraux* qui sont dotés de caractéristiques métaphysiques centrales pour la responsabilité morale, telles que la conscience, l'intentionnalité et le libre arbitre (Moor, 2009).

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

Plusieurs auteurs comme Martin Gibert et Dominic Martin distinguent la *patience morale* et l'*agentivité morale* (Gibert et Martin, 2022). Cette distinction n'est pas seulement faite en éthique de l'IA. En général, on considère que les agents moraux ont des responsabilités morales, tandis que les patients moraux ont des droits, car ils ont des intérêts qui peuvent être lésés et on peut donc leur causer du tort (Singer, 2015). Or, comme le note Vincent Müller, il semble évident que « certaines entités sont des patients sans être des agents, par exemple de simples animaux qui peuvent ressentir la douleur, mais ne peuvent pas faire des choix justifiés. D'autre part, il est généralement admis que tous les agents moraux sont également des patients moraux, dans une perspective kantienne par exemple » (Müller, 2021 ; je traduis). Cette distinction est importante dans le cadre de notre propos. En effet, comme le précise Véliz elle-même, son argument ne concerne pas la patience morale, mais plutôt l'agir moral, donc l'agentivité morale.

Analysons maintenant cet argument : P [1] La sentience est une condition nécessaire pour être un agent moral. Pourquoi la sentience plutôt que la conscience comme condition nécessaire à l'agentivité morale ? Véliz le justifie par la possibilité qu'il y ait des agents conscients, mais qui n'ont pas d'expériences de plaisir ou de douleur, ou à qui il manquerait les émotions associées à ces expériences. Avoir des expériences typiquement sentientes est ce qui compte le plus pour l'agentivité morale (Véliz, 2021). Son argument s'inscrit dans une tradition humienne qui met l'accent sur le rôle des émotions morales dans l'évaluation de l'agir moral (Rietti, 2003 ; Tappolet, 2016).

En effet, la disposition à avoir des sentiments, des émotions ou à être empathique semble jouer un rôle déterminant dans l'agir moral. N'est-ce pas justement cette capacité qui informe et motive nos actions morales ? Pour reprendre une formule de David Shoemaker : « sans la capacité de ressentir, nous serions (par définition) dépourvus de la capacité de nous soucier des autres [...] Perdre sa capacité à se soucier des autres, c'est perdre son identité en tant qu'agent cohérent, [car] les émotions que nous ressentons font de nous les agents que nous sommes » (Shoemaker, 2003, p. 114 ; je traduis). Dès lors, comment une entité qui ne ressent rien pourrait-elle être qualifiée d'agent moral authentique ?

Voyons le second argument : P [2] Les zombies moraux et les algorithmes ne sont pas sentients. Selon Véliz, la sentience est absente chez les zombies (moraux) et les algorithmes. Pour motiver cette prémisse, la philosophe s'emploie avant tout à légitimer l'application de l'expérience de pensée des zombies dans le domaine de la moralité.

Rappelons d'abord avec Robert Kirk que traditionnellement, en philosophie de l'esprit, « les zombies sont des créatures imaginaires conçues pour éclairer les problèmes liés à la conscience et à sa relation avec le monde physique. Contrairement à ceux des films ou de la sorcellerie, ils sont exactement comme nous dans tous les aspects physiques, mais sans expérience consciente » (Kirk, 2023 ; je traduis). Or, explique Véliz (2021), l'expérience de pensée des zombies philosophiques peut être transposée dans le champ de la morale, car une analogie est possible avec la sentience. La philosophe propose qu'on imagine, à l'instar des zombies philosophiques, des zombies spécifiquement moraux. L'idée est que les zombies moraux seraient des entités qui agissent à tout point de vue comme des agents moraux humains. Leurs

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

comportements moraux seraient impossibles à différencier de ceux des agents humains moraux. Je précise ici que l'autrice a en tête une sorte d'indiscernabilité de comportement que je qualifierais de typique (comportement d'un humain standard et, précisément ici, d'un agent moral standard), contrairement à une indiscernabilité stricte (qui représenterait les doubles identiques et fonctionnels d'un individu déterminé et précis) (Esfeld, 2017).

Comme les agents moraux humains en général, les zombies moraux seront engagés comportementalement et activement dans des situations moralement significatives. Ils sont comme les zombies philosophiques traditionnels, mais impliqués essentiellement dans les comportements moraux. Ils agissent bien ou mal. Ils peuvent ainsi mentir, reproduire des inégalités sociales, mais aussi lutter pour la justice ou contre le réchauffement climatique. Il n'en reste pas moins que, contrairement aux agents moraux humains, ces zombies moraux « ne ressentiraient ni douleur, ni plaisir, ni culpabilité, ni honte, ni aucune autre émotion morale » (Véliz, 2021, p. 2 ; je traduis).

L'idée centrale de Véliz est que les algorithmes seraient des équivalents de ces zombies moraux puisqu'ils sont fonctionnellement identiques. Il est quasi certain, dans l'état actuel de nos connaissances, que nos algorithmes ne sont pas sentients. Ils sont ainsi, du point de vue moral, « incohérents en tant qu'agents moraux, car ils n'ont pas la compréhension morale nécessaire pour être moralement responsable » (Véliz, 2021, p. 2 ; je traduis). En l'absence de test de conscience fiable, nous n'avons actuellement aucun moyen de savoir si ces agents ont la capacité de ressentir de la douleur ou du plaisir. Or, il est certain que la connaissance expérientielle consciente de la douleur ou du plaisir est essentielle pour agir comme un agent moral qui a une compréhension morale nécessaire de ses actions (Véliz, 2016).

La conclusion : C [3] Donc, les zombies moraux et les algorithmes ne sont pas des agents moraux découle logiquement des prémisses. L'argument est valide. Faisons néanmoins quelques remarques.

Même si la conclusion semble radicale et forte, précisons que c'est sous réserve d'avoir la certitude que les algorithmes ne sont pas sentients. Le problème épistémique relatif à la conscience demeure entier. Comment savoir avec certitude si les robots sont sentients ? Dans son texte de 2016, Véliz admettait déjà cette difficulté liée au fait que nous n'avons pas le moyen de déterminer véritablement si ces agents artificiels ont la capacité de ressentir de la douleur ou du plaisir (Véliz, 2016). C'est pourquoi, s'il arrive que nous ayons des machines qui sont hautement semblables à nous, en l'absence de test de conscience certain, la prudence devrait être de mise (Véliz, 2021).

2. Objection à la P [1]

Dans cette section, je vais essayer de remettre en question la véracité de la prémisse [1]. En effet, on peut contester le fait que la sentience soit une condition nécessaire à l'agentivité morale. Je tenterai de montrer que les agents collectifs sont un cas d'agentivité morale sans sentience. Or, si en tant qu'entité agentielle sans conscience phénoménale, ils peuvent être des agents moraux, alors la sentience ne peut pas être un critère nécessaire à l'agentivité morale. Une entité dépourvue de conscience ne peut, par définition, être sentiente. En revanche, l'inverse pourrait faire l'objet d'un

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

débat. Autrement dit, on pourrait se demander si une entité sentiente est nécessairement consciente.

Véliz estime que sa conception de l'agentivité morale est à l'opposé de celle de Luciano Floridi et J. W. Sanders (Floridi et Sanders, 2004). Ces derniers soutiennent qu'il est concevable d'avoir une agentivité morale sans esprit, ni libre arbitre, ni émotions : ils envisagent une moralité dépourvue de toute forme de conscience (« mind-less morality »). Pour les besoins de mon argumentation, j'adopte ici la définition d'*agent* de Floridi et Sanders (2004). Elle me permet de montrer par la suite que la définition de Véliz exclut certains agents moraux pourtant évidents. Une entité peut être considérée comme un agent s'il s'agit d'un système qui est (a) *interactif*, c'est-à-dire capable d'avoir une action et une rétroaction avec son environnement : le système et son milieu agissent l'un sur l'autre par des mécanismes d'entrée et de sortie (input et output). Il doit également être (b) *autonome*, c'est-à-dire avoir la capacité de modifier son état sans interaction directe avec son environnement ; autrement dit, une entité qui peut faire des transitions pour changer son état (ce qui suppose qu'un agent devrait minimalement avoir deux états). Enfin, le système doit être (c) *adaptable*, c'est-à-dire que les interactions de l'entité peuvent changer les règles de la transition par laquelle elle change d'état (Floridi et Sanders, 2004).

C'est en ayant à l'esprit ces considérations que je formulerai mon objection. Elle peut se résumer comme suit : il est possible d'avoir des agents moraux non sentients, comme le montre le cas des agents collectifs. En effet, les organisations ou les entreprises, par exemple, peuvent remplir les conditions de l'agentivité (a), (b) et (c), dans la mesure où « une société ou un organisme de gestion ou de direction, à un niveau d'agentivité approprié, interagit avec ses employés, ses sous-structures constitutives et d'autres organisations ; elle est capable de procéder à des changements d'états déterminés en interne et d'adapter ses stratégies de prise de décision et, partant, d'action » (Floridi et Sanders, 2004, p. 14 ; je traduis). Or, une telle agentivité semble ne pas nécessiter la possession d'états mentaux. La dimension« morale » de cette agentivité sera explicitée plus loin. Mais procédons par étape. D'abord, peut-on avoir une agentivité pertinente sans esprit, ou sans la présence d'un cerveau ?

2.1. Agentivité et états mentaux

Le concept d'agentivité, en philosophie du moins, a souvent été utilisé au sens circonscrit d'actions intentionnelles. On peut définir le concept d'intention en lien avec l'agentivité en suivant Deborah Perron Tollefsen : « les intentions sont ce qui distingue les actions — les choses que je fais — des simples événements — les choses qui m'arrivent » (Tollefsen, 2015, p. 27 ; je traduis). Nous clignons des yeux et nous trébuchons, par exemple, sans en avoir l'intention. Nous posons d'autres gestes de manière intentionnelle, comme le fait que je sois en train de taper ce texte en ce moment en choisissant mes mots. C'est l'intentionnalité qui me confère le pouvoir de contrôle sur les actions. En ce qui nous concerne, nous donnons la définition suivante de l'intentionnalité : c'est la capacité d'une entité à agir intentionnellement. Précisons tout de suite qu'elle n'est pas à confondre avec l'intentionnalité entendue comme « le pouvoir qu'ont les esprits et les états mentaux de se rapporter à des choses, des propriétés et des états de choses, ou de les représenter. Dire des états mentaux d'un

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

individu qu'ils ont une intentionnalité revient à dire qu'ils sont des représentations mentales ou qu'ils ont un contenu » (Jacob, 2023 ; je traduis).

Il s'ensuit qu'avoir une agentivité, en ce sens, c'est être une entité capable d'agir. Une entité exerce son agentivité dans la mesure où elle a la capacité d'agir intentionnellement dans certaines situations. Un agent peut également poser des gestes non intentionnellement. La distinction découle de ce que dans un cas, il y a un contenu représentationnel qui accompagne l'acte, alors que dans l'autre, l'acte en est dépourvu. Cette conception remonte en particulier aux travaux d'Elizabeth Anscombe (1957), ainsi qu'à ceux de Donald Davidson (1973), qui soutiennent tous deux que 1) la capacité d'agir d'un agent doit être essentiellement liée à la capacité d'exécuter des actions avec une certaine intentionnalité, et que 2) l'intentionnalité est plus fondamentale que l'action elle-même dans la mesure où elle la constitue.

L'intentionnalité est par ailleurs étroitement liée à l'idée d'*agir pour des raisons* (Schlosser, 2019). Une telle approche, qui propose de rendre compte de l'agentivité en se basant sur l'intentionnalité, fait appel à des représentations mentales. Or, les représentations mentales peuvent avoir des contenus comme la croyance, le désir, la peur. Mais « les croyances sont des attitudes propositionnelles [...] qui impliquent un sujet, un contenu et une attitude » (Tollefsen, 2015, p. 8-9 ; je traduis). Les croyances peuvent ainsi posséder une condition de vérité dans la mesure où elles se rapportent à un état de fait dans le monde doté ou non d'un contenu représentationnel vrai ou faux. Les attitudes propositionnelles supposent une attitude au regard d'une proposition se référant à un état du monde (Tollefsen, 2015, p. 9-10 ; je traduis).

Ainsi, les états mentaux comme les croyances, les souvenirs ou les désirs sont ce que les philosophes appellent habituellement des états intentionnels. Ce sont aussi des états qui semblent conscients, c'est-à-dire qu'ils auraient des « *qualia* » et seraient constitutifs d'une conscience phénoménale. Autrement dit, il peut y avoir un effet à les ressentir. Ainsi, cette conception de l'agentivité semble mal prendre en compte, ou de manière insuffisante, des formes d'agentivité non humaines tout aussi significatives, mais dépourvues d'états mentaux représentationnels liés à la conscience. En effet, certaines entités telles que les entreprises, les collectivités et les groupes peuvent être perçues *prima facie* comme dépourvues de cerveau ou d'esprit et par conséquent d'états mentaux. Cependant, nous les considérons généralement comme capables d'agir moralement et d'assumer des responsabilités morales. Dans les prochains paragraphes, nous examinerons spécifiquement le cas des groupes en tant qu'agents.

Considérons d'abord les agents non humains sommaires avant d'examiner des formes d'agentivité humaines.

Prenons l'exemple d'une entité non humaine dépourvue d'états mentaux représentationnels. Pour Xabier Barandiaran et ses collègues, des systèmes protocellulaires comme les bactéries peuvent représenter un cas paradigmatique d'agentivité minimale authentique. L'agentivité d'un système ou d'une entité peut être définie comme « une organisation autonome qui régule de manière adaptative son couplage avec son environnement ainsi que sa propre survie » (Barandiaran et al., 2009, p. 1 ; je traduis).

On le voit, cette définition comporte les trois conditions — (a) interactif, (b) autonome et (c) adaptable — de l'agentivité telle que définie par Floridi et Sanders (2004). Ainsi, ces conditions s'appliquent en réalité aussi bien à une agentivité de

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

groupe qu'à une agentivité sommaire. Cela nous indique que l'agentivité pourrait être mieux saisie par degrés et dans un continuum, notamment pour les systèmes d'IA (Longin, 2020), un agent pouvant être plus ou moins interactif, autonome et adaptable.

Comment analyser les actions provenant d'une source humaine non intentionnelle ? Regardons du côté de la tradition phénoménologique, de Husserl à Merleau-Ponty en passant par Heidegger. Ces auteurs mettent en lumière l'engagement du corps dans son environnement. Ils s'intéressent à des habiletés et des dispositions qui, la plupart du temps, sont des actions réflexes et inconscientes. L'agent, explique Markus Schlosser, a un « engagement habile et en "temps réel" avec le monde : la capacité de s'engager avec les autres et avec sa propre situation en répondant aux exigences de la situation d'une manière habile et souvent sans effort, sans délibération consciente, raisonnement ou planification » (Schlosser, 2019 ; je traduis). Merleau-Ponty donne l'exemple de la conduite automobile, ou encore de la personne portant un chapeau dont les mouvements anticipent inconsciemment l'évitement de la porte (Merleau-Ponty, 1945). On pense aussi à l'improvisation en jazz. Comme le donnent à voir certaines versions des approches « enactivistes » (enactive approach) (Chemero, 2009 ; Hutto et Myin, 2012), ce type d'agentivité sans représentation mentale nous offre des raisons de croire que toute manifestation d'agentivité, même humaine, pourrait et devrait être expliquée sans l'attribution d'états mentaux représentationnels (Schlosser, 2019).

S'il est possible de déterminer des formes d'agentivité qui ne nécessitent pas l'usage conscient d'états mentaux, nous pouvons maintenant en venir à l'agentivité de groupe. Quelles peuvent être les conditions de possibilité d'une agentivité de groupe, autrement dit d'un groupe considéré comme un agent à part entière ? Un groupe, bien qu'il soit constitué d'individus, n'a pas en lui-même un cerveau ou un esprit. Peut-on néanmoins lui attribuer une agentivité intentionnelle ?

2.2. Groupes comme agents intentionnels

Dans cette section, je m'appuierai sur l'interprétativisme, une théorie défendue par de nombreux philosophes (Kriegel, 2011a ; 2011b ; Dennett, 1987 ; 1991 ; Haugeland, 1990 ; Davidson, 1973). Selon l'interprétativisme, « si nous réussissons à donner un sens à un autre être [successfully make sense of another being], c'est-à-dire à comprendre et à interpréter son comportement en utilisant notre psychologie populaire, il s'agit d'un agent intentionnel » (Tollefsen, 2015, p. 97 ; je traduis). Ma stratégie argumentative consiste à dire que même si l'on accepte l'hypothèse selon laquelle l'agentivité repose sur une forme d'intentionnalité, c'est l'interprétativisme qui offre les meilleurs outils conceptuels pour saisir la réalité de l'agentivité. L'intérêt d'une telle approche est qu'elle évite de se perdre dans des discussions métaphysiques compliquées sur la nature des phénomènes mentaux.

On peut alors se concentrer sur une question non moins fondamentale : quelles sont les caractéristiques essentielles permettant d'interpréter efficacement un comportement et de l'attribuer avec succès à un agent ? Il s'agit ainsi d'adopter ce que Daniel Dennett (1987) a appelé une « posture intentionnelle » (intentional stance). Que veut dire adopter une posture intentionnelle envers une entité ? C'est essayer de

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

prédire et d'expliquer le comportement de l'entité comme s'il s'agissait d'un agent rationnel dont les actions sont dirigées par ses intentions, croyances et désirs.

C'est justement cette posture qui, selon moi, légitime l'attribution d'une agentivité intentionnelle à un système ou à une entité dont le fonctionnement nous porte à croire que ces actions sont orientées par une certaine attitude propositionnelle, même si la source de l'action ne semble pas posséder d'états phénoménaux. Uriah Kriegel (2011a ; 2011b) exprime bien cette idée, qui est reprise par David Bourget et Angela Mendelovici :

> un état non phénoménal 'S' a un certain contenu intentionnel dérivé 'C', dans la mesure où un interprète idéal est disposé à attribuer 'C' à 'S'. Un interprète idéal est un être parfaitement rationnel qui connaît tous les faits phénoménaux et non phénoménaux (mais pas les faits intentionnels dérivés) dans le monde. Selon le point de vue dérivativiste qui en découle, les états intentionnels non phénoménaux dérivent des états intentionnels phénoménaux de l'interprète idéal. (Bourget et Mendelovici, 2019 ; je traduis)

Lorsque nous prenons la posture intentionnelle, nous ne cherchons *pas simplement* (je souligne) à expliquer ou à prédire une action ou un énoncé de l'agent ou du système, mais nous visons l'agent en lui-même, c'est-à-dire que « le sujet auquel on attribue l'état intentionnel est l'ensemble du système, ou l'ensemble de la personne » (Tollefsen, 2015, p. 98 ; je traduis). Lorsque je prête à mon chien l'intention de sortir pour une promenade parce qu'il se met devant la porte aux environs de 16 heures, je n'attribue pas ce désir à son cerveau ou à un quelconque système nerveux. C'est bien à lui, en tant qu'individu, que j'attribue l'envie d'une promenade. De même, lorsque j'attribue une croyance ou des intentions à mes parents ou à mon enfant, encore une fois, je ne mets pas ces attitudes propositionnelles sur le compte de leur cerveau ou de quelque esprit en eux. Bien évidemment, je me doute qu'ils ont un cerveau. Mais ce qu'il importe de comprendre ici, c'est qu'il n'est pas nécessaire de faire appel à un cerveau ou à un esprit pour attribuer des états intentionnels à un système, faisant de ce dernier un agent intentionnel.

Autrement dit, ce que semble indiquer une réflexion sérieuse sur nos pratiques, c'est qu'il n'est pas nécessaire d'avoir un cerveau ou un esprit pour avoir un état mental. En effet, les états mentaux ne sont pas des états internes du cerveau, mais des états d'un agent ou d'un système, puisqu'il s'agit d'une attribution indirecte par dérivation, comme le montre Kriegel (Kriegel, 2011a ; 2011b).

Si ce qui précède est vrai, il nous est donc possible d'attribuer une agentivité intentionnelle à un groupe en dépit du fait qu'il s'agit d'un système dépourvu d'esprit. Si nous adoptons la posture intentionnelle envers un groupe, nous pouvons lui prêter, d'une manière holistique, c'est-à-dire en tant que système ou entité agentiels, des états intentionnels. Nous supposons alors que le groupe partage des normes rationnelles et a une perspective rationnelle unifiée, de sorte que nous pouvons lui attribuer des croyances, des intentions ou des désirs. Cela n'implique pas que nous prenions la posture intentionnelle envers tous les groupes. Et encore moins envers toute entité comme la foudre ou l'ouragan, par exemple.

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

Ainsi, lorsque nous affirmons que le Département de philosophie de l'Université de Montréal *désire* attirer de bons chercheurs, nous adoptons une posture intentionnelle envers le département comme groupe qui, en tant que tel, a ce désir. Cette stratégie, me semble-t-il, nous évite plusieurs écueils : d'abord, je le répète, celui de nous perdre dans des discussions métaphysiques qui peuvent parfois être obscures, voire insolubles, par exemple sur la nature des intentions. Ensuite, elle nous permet de contourner le problème épistémologique complexe de la conscience, c'est-à-dire comment avoir la certitude d'un accès à la conscience d'autrui. Le problème des autres esprits (other minds problem) est l'un des plus difficiles à résoudre. Enfin, cette stratégie maintient l'observateur rationnel comme juge impartial, comme l'indique l'approche dérivativiste mentionnée plus tôt (Kriegel, 2011a ; 2011b). Cela devrait nous éviter de tomber à la fois dans un anthropocentrisme suspect (considérer uniquement les humains comme des agents intentionnels) et dans un libéralisme excessif ou imprudent (considérer toute entité comme un agent intentionnel) quant à l'attribution d'agentivité à des systèmes ou des entités divers et variés.

De ce point de vue, il me semble que l'intentionnalité est un critère qui vaut bien la sentience pour déterminer l'agentivité en tant que telle et, *a fortiori*, l'agentivité morale. Or, il semble possible d'attribuer une certaine intentionnalité à des systèmes algorithmiques complexes, des agents qui agissent comme le ferait un agent moral humain ayant des intentions. Il convient alors de se demander si, dans ces conditions, la conscience est encore nécessaire à l'agentivité morale.

2.3. Un agent dépourvu de conscience phénoménale peut-il être moral ?

Nous venons de voir que l'attribution d'une agentivité intentionnelle à des entités ou des systèmes est plausible, même si ceux-ci sont dépourvus d'esprit. En réalité, la possession d'un cerveau ou d'un esprit ne semble pas nécessaire pour attribuer des états mentaux à des entités. Nous pouvons leur conférer une certaine agentivité intentionnelle. Comme nous l'affirmions plus haut, les états mentaux sont à la fois des états intentionnels et des états avec des contenus qualitatifs de conscience phénoménale.

Or, il n'est pas certain que la conscience phénoménale soit nécessaire pour avoir des attitudes propositionnelles. En effet, les états qualitatifs ne supposent pas forcément une intentionnalité, un but, une croyance ou un espoir. Nous pouvons, par exemple, attribuer des désirs au Département de philosophie de l'Université de Montréal. Si je peux ressentir l'effet que procure le fait de regarder du rouge, ou le ciel bleu de la Côte d'Azur, l'existence d'un effet lié au fait d'*avoir l'intention* d'aller visiter les plages du sud de la France, ou de *croire* faussement que Berlin est la capitale de l'Argentine, ou encore d'*espérer* qu'il fasse beau demain, paraît plus douteuse.

Une manière simple de procéder à l'analyse des relations possibles entre conscience et attitudes propositionnelles serait de les séparer, tout en distinguant la conscience phénoménale de la conscience d'accès, comme le fait Ned Block. Il existe des preuves concrètes provenant d'expérimentations qui montrent que le système responsable des états qualitatifs de la conscience ne correspond pas aux états impliqués dans l'accès cognitif aux états du monde, comme le donne à penser la théorie de l'espace de travail global, par exemple. C'est pourquoi, selon Block (2008), il est nécessaire de faire une

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

distinction entre la conscience phénoménale et la conscience d'accès, cette dernière relevant davantage du domaine cognitif. La conscience d'accès nous permet une gestion globale du système cognitif. Cette distinction permet à Block de soutenir que la conscience d'accès est nécessaire à l'agentivité intentionnelle, tandis que la conscience phénoménale ne semble pas essentielle pour les attitudes propositionnelles (Tollefsen, 2015).

Une telle distinction est importante et intéressante, car elle recoupe et conforte notre définition de l'agentivité à partir des notions d'états intentionnels, d'interaction et d'adaptation. Ainsi, il semble possible d'imaginer des entités telles que des systèmes d'intelligence artificielle ou des robots qui ont une conscience d'accès et des états intentionnels, mais qui sont dépourvues de toute conscience phénoménale : elles ne ressentent ni douleur, ni peur, ni joie, mais elles ont la capacité de représenter un état du monde qui peut être vrai ou faux. De fait, cela autoriserait à penser que la conscience phénoménale n'est pas nécessaire à l'agentivité (Tollefsen, 2015, p. 52). À partir de là, j'avance, à la suite de Tollefsen et d'autres philosophes, que les attitudes propositionnelles, qui sont des états mentaux ou intentionnels ou encore représentationnels, ne nécessitent pas de conscience phénoménale. Je soutiendrai donc que les groupes n'ont pas besoin d'être phénoménalement conscients pour être des agents, et en l'occurrence des agents moraux. Il suffit qu'ils puissent avoir des états intentionnels, dont plusieurs sont des attitudes propositionnelles, pour posséder une capacité agentielle suffisamment pertinente pour l'action morale.

Les groupes en tant qu'agents peuvent avoir des comportements moraux, ce qui permet à un sujet qui adopte la posture intentionnelle de leur attribuer une agentivité intentionnellement morale. En effet, si nous reprenons l'intentionnalité comme étant liée à l'idée d'agir pour des raisons (Schlosser, 2019), l'agentivité morale d'entités qui ne possèdent pas nécessairement une conscience phénoménale devient aisément concevable, car nous pouvons déterminer si celles-ci agissent conformément ou non à des raisons morales. Nous comprenons mieux la légitimité de les considérer comme des agents moraux en dépit de l'absence de toute conscience phénoménale ou d'états mentaux. Les théories normatives rationalistes, comme le déontologisme kantien ou le conséquentialisme, semblent très bien nous permettre d'évaluer la moralité des actions de telles entités.

Il s'ensuit qu'agir conformément à un devoir moral ou à une raison morale serait suffisant, tant qu'un observateur rationnel peut interpréter et évaluer cette action en tant que telle. Nous pouvons par exemple estimer que l'Université X ou l'entreprise Y ou l'État Z se sont rendus moralement coupables d'actions qui contreviennent à certains devoirs moraux. Il en va de même pour des actions dont les conséquences sont moralement répréhensibles, si l'on adopte une perspective conséquentialiste : par exemple, financer des projets ayant un impact écocidaire, soutenir directement ou indirectement des crimes contre l'humanité, ou encore promouvoir la souffrance d'êtres sentients par une exploitation injuste.

En résumé, nous avons de bonnes raisons de penser que la conscience n'est pas nécessaire à l'agentivité intentionnelle. Par ailleurs, il est par définition impossible d'imaginer une entité sentiente sans conscience. Une entité non consciente ne semble pas pouvoir être sentiente. Donc, si une entité dépourvue de conscience peut avoir une agentivité intentionnelle morale, cela signifie qu'il n'est pas nécessaire d'être sentient

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

pour être un agent moral. Mais certains pourraient avoir des réticences à adopter une telle perspective, qui peut évidemment susciter certaines objections.

3. Une objection possible à l'argument de l'agentivité sans états mentaux, ni conscience phénoménale

L'objection que je présente ici relève plus d'une querelle de famille et n'est pas radicalement antagonique à mon argument de la section 2. En effet, certains auteurs admettent l'idée d'une agentivité (morale) de groupe, mais estiment qu'elle ne peut que reposer ultimement sur des individus humains.

Une objection plausible serait qu'une agentivité de groupe ne peut être conçue qu'en termes d'agentivité partagée (shared agency). Dans le cas des collectifs ou des groupes, on parle en général d'une agentivité partagée ou d'une agentivité collective. L'agentivité *partagée* fait référence à une action effectuée par au moins deux individus : un orchestre qui exécute un morceau, ou deux personnes qui poussent une voiture embourbée. L'agentivité *collective*, pour sa part, indique qu'au moins deux personnes agissent en tant que groupe et non en tant qu'individus : elles en respectent les procédures et sont en accord avec les principes qui organisent et constituent le groupe en tant que tel (Schlosser, 2019). Un département de philosophie pourrait en être un exemple. Autrement dit, hors des procédures et des principes d'action qui constituent le groupe, il ne peut être question d'une agentivité collective puisque cette sorte d'agentivité n'est possible qu'en tant qu'elle opère par une structuration qui tient au-delà des individus particuliers qui la composent et qui n'en sont qu'une partie du mécanisme agentiel. L'agentivité partagée et l'agentivité collective se distinguent de l'agentivité qui résulte de l'assemblage d'individus agissant individuellement par le fait qu'il y a, pour les premières, une intentionnalité commune, un « nous intentionnel » (we-intention), ce qui n'est pas le cas de la seconde (Roth, 2017).

La question ici est de savoir sous quelles conditions les groupes peuvent avoir des croyances, des intentions, des souhaits ou des souvenirs, lesquels sont tous des attitudes propositionnelles. Certains auteurs soutiennent une conception dite sommative (agrégative/additionnelle), c'est-à-dire qui pose qu'un groupe a des croyances si et seulement si la majorité des *individus* membres du groupe partagent ces croyances. Anthony Quinton exprime bien ce qu'il faut comprendre de cette conception.

> Bien entendu, nous parlons librement des propriétés mentales et de l'action d'un groupe comme nous le faisons pour des personnes individuelles. On dit que les groupes ont des croyances, des émotions et des attitudes, qu'ils prennent des décisions et font des promesses. Mais ces façons de parler sont clairement métaphoriques. Attribuer des prédicats mentaux à un groupe est toujours une manière indirecte d'attribuer de tels prédicats à ses membres. Dans le cas d'états mentaux tels que les croyances et les attitudes, les attributions sont, ainsi que je l'ai appelé, de type additionnel. (Quinton, 1976, p. 17 ; je traduis)

Cette conception peut être résumée formellement ainsi : pour tout groupe (A), (A) croit que X (X étant une proposition), si et seulement si l'ensemble (A) ou la majorité des membres de (A) croient que X.

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

Cette conception présente différentes difficultés, comme la manière d'attribuer une croyance ou une intention à un groupe dont chaque membre aurait la même idée ou croyance sans la communiquer aux autres membres. Il peut également exister des groupes dans lesquels une minorité non légitime impose ses croyances à l'ensemble du groupe du fait que les autres membres sont timides ou n'ont pas véritablement voix au chapitre. Peut-on vraiment parler dans ces cas de croyances de groupe ? Il semble difficile d'adopter cette conception sans y ajouter d'autres conditions dans lesquelles il peut être intéressant d'additionner les croyances des individus d'un groupe. Même si l'on arrive à le faire, cela ne dit rien non plus sur le groupe en tant qu'entité à part entière. Il s'agit plutôt d'une conception d'agentivité partagée, dans la mesure où les individus partagent des croyances, des intentions, etc., qui ne rendent pas compte des croyances d'un groupe en tant que tel (Tollefsen, 2015).

Une autre approche propose de se baser sur la légitimité de certains membres du groupe, même s'ils ne sont pas majoritaires, et d'accepter leurs croyances, leurs intentions, voire leurs jugements, comme étant ceux du groupe. Pour ce faire, Raimo Tuomela introduit la notion d'agent « opérant » versus « non opérant », dans la mesure où les agents opérants occupent une position spéciale qui leur permet d'avoir une certaine vision à laquelle les autres membres s'associent ou se joignent (Tuomola, 1992). Selon Tuomela, en s'appuyant sur la notion d'action collective, il est possible de proposer une conception de la croyance de groupe qui échappe au problème de l'approche sommative ou additionnelle (Tuomela, 1992).

En effet, tous les membres n'ont pas besoin de jouer un rôle dans la détermination de la croyance du groupe. Mais dans ces conditions, que signifie concrètement cette croyance dans un groupe où seulement une ou deux personnes sont autorisées et parviennent à imposer leur vision à l'ensemble des membres ? Il semble que les états mentaux d'un tel ordre soient extrinsèques et non propres à l'agentivité humaine individuelle. L'agentivité de groupe devrait alors être comprise comme une agentivité fonctionnelle (Tuomela, 2013). Michael Bratman considère par exemple que l'intentionnalité partagée devrait se concevoir « comme un état de choses consistant principalement en des attitudes appropriées de chaque participant individuel et leurs interrelations » (Bratman, 1993, p. 99-100 ; je traduis). Il n'en demeure pas moins qu'aucune de ces approches ne propose une conception des différentes caractéristiques de l'agentivité qui ne soit pas, d'une manière ou d'une autre, réductible aux individus d'un groupe et non au groupe lui-même comme agent. Selon elles, le groupe en tant que tel ne peut en aucun cas avoir des états intentionnels. Quand bien même certains aspects de l'analogie entre les groupes humains et les systèmes d'IA semblent contestables, ceux-ci n'en ont pas moins en commun de pouvoir être interprétés comme ayant des états intentionnels.

Évaluation conclusive

Rappelons l'argument de Véliz :

P [1] La sentience est une condition nécessaire pour être un agent moral.

P [2] Les zombies moraux et les algorithmes ne sont pas sentients.

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

C [3] Donc, les zombies moraux et les algorithmes ne sont pas des agents moraux.

Dans cet article, j'ai contesté l'argument en niant la prémisse [1]. J'ai procédé en soutenant qu'il existe des entités comme les groupes qui sont des agents intentionnels sans être sentientes et, plus radicalement, sans avoir de conscience phénoménale. L'intentionnalité étant déterminante dans l'évaluation de l'agir moral, cela nous indique que ces systèmes ou entités intentionnels, sans conscience ni sentience, pourraient être des agents moraux (comme une armée, une nation, etc.). Il s'ensuit que la sentience n'est pas requise pour l'agentivité morale. Certains pourraient tout de même objecter qu'une telle agentivité n'est possible qu'en tant qu'elle est réductible à des agentivités morales humaines qui sont partagées ou additionnées. Il semble au contraire que si nous adoptons une approche interprétativiste, nous évitons un tel anthropocentrisme tout en prenant garde de ne pas voir dans toute entité une source d'agentivité intentionnelle et morale. C'est la raison pour laquelle j'estime que cette objection possible à mon argument de la section 2 n'est pas fatale. Pour les critères de l'agentivité morale, l'intentionnalité, vue dans une perspective interprétativiste, me semble suffisante et la sentience n'est pas nécessaire (comme le montre l'agentivité morale des groupes). Donc, en l'état, rien n'interdit d'avancer que les algorithmes sont des agents moraux. En définitive, si l'on accepte que certains groupes puissent être des agents moraux, on a de bonnes raisons de dire que c'est aussi le cas de certains algorithmes qui partagent avec les groupes le fait d'être considérés comme intentionnels sans être sentients.

Conflits d'intérêts. L'auteur n'en déclare aucun.

Références bibliographiques

Anderson, M. et S. L. Anderson (2011). *Machine Ethics*. Cambridge University Press.

Anscombe, G. E. M. (1957). Intention. *Proceedings of the Aristotelian Society*, *57*(1), 321–332. https://doi.org/10.1093/aristotelian/57.1.321

Barandiaran, X. E., Di Paolo, E. et M. Rohde (2009). Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action. *Adaptive Behavior*, *17*(5), 367–386. https://doi.org/10.1177/1059712309343819

Block, N. (2008). Phenomenal and Access Consciousness. *Proceedings of the Aristotelian Society*, *108*(1), 289–317. https://doi.org/10.1111/j.1467-9264.2008.00247.x

Bourget, D. et A. Mendelovici (2019). Phenomenal Intentionality. Dans E. N. Zalta (dir.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/phenomenal-intentionality/

Bratman, M. E. (1993). Shared Intention. *Ethics*, *104*(1), 97–113. https://doi.org/10.1086/293577

Chemero, A. P. (2009). Radical Embodied Cognitive Science. *MIT Press*. https://doi.org/10.7551/mitpress/8367.001.0001

Davidson, D. (1973). Radical Interpretation. *Dialectica*, *27*(3-4), 313–328. https://doi.org/10.1111/j.1746-8361.1973.tb00623.x

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

Dennett, D. C. (1991). *Consciousness Explained*. Little Brown.

Esfeld, M. (2017). Philosophie de l'esprit (A). Dans M. Kristanek (dir.), *L'encyclopédie philosophique*. https://encyclo-philo.fr/philosophie-de-lesprit-a

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

Floridi, L. et J. W. Sanders (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/b:mind.0000035461.63578.9d

Formosa, P. et M. Ryan (2021). Making Moral Machines: Why We Need Artificial Moral Agents. *AI & Society*, *36*, 839–851. https://doi.org/10.1007/s00146-020-01089-6

Gibert, M. et D. Martin (2022). In Search of the Moral Status of AI: Why Sentience Is a Strong Argument. *AI & Society*, *37*(1), 319–330. https://doi.org/10.1007/s00146-021-01179-z

Haugeland, J. (1990). The Intentionality All-Stars. *Philosophical Perspectives*, *4*, 383–427. https://doi.org/10.2307/2214199

Hutto, D. D. et E. Myin (2012). *Radicalizing Enactivism: Basic Minds without Content*. MIT Press. https://doi.org/10.7551/mitpress/9780262018548.001.0001

Jacob, P. (2023). Intentionality. Dans E. N. Zalta et U. Nodelman (dir.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2023/entries/intentionality/

Kirk, R. (2023). Zombies. Dans E. N. Zalta et U. Nodelman (dir.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/zombies/

Kriegel, U. (2011a). *The Sources of Intentionality*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199742974.001.0001

Kriegel, U. (2011b). Cognitive Phenomenology as the Basis of Unconscious Content. Dans T. Bayne et M. Montague (dir.), *Cognitive Phenomenology* (p. 79-102). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199579938.001.0001

Longin, L. (2020). Towards a Middle-Ground Theory of Agency for Artificial Intelligence. Dans M. Nørskov, J. Seibt et O. Quick (dir.), *Culturally Sustainable Social Robotics: Proceedings of Robophilosophy 2020* (p. 17–26). IOS Press. https://doi.org/10.3233/faia200897

Moor, J. (2009). Four Kinds of Ethical Robots. *Philosophy Now, 72*, 12–14.

Müller, V. C. (2021). Ethics of Artificial Intelligence and Robotics. Dans E. N. Zalta (dir.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/

Merleau-Ponty, M. (1945). *La phénoménologie de la perception*. Gallimard.

Quinton, A. (1976). The Presidential Address: Social Objects. *Proceedings of the Aristotelian Society*, *76*(1), 1–28. https://doi.org/10.1093/aristotelian/76.1.1

Rietti, S. (2003). *Role of Emotion in Moral Agency: Some Meta-Ethical Issues in the Moral Psychology of Emotion* [thèse de doctorat]. University of Edinburgh. https://era.ed.ac.uk/handle/1842/4051

Roth, A. S. (2017). Shared Agency. Dans E. N. Zalta (dir.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2017/entries/shared-agency/

Schlosser, M. (2019). Agency. Dans E. N. Zalta (dir.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/

Sharkey, A. (2020). Can We Program or Train Robots to Be Good? *Ethics and Information Technology*, *22*(4), 283–295. https://doi.org/10.1007/s10676-017-9425-5

Shoemaker, D. W. (2003). Caring, Identification, and Agency. *Ethics*, *114*(1), 88–118. https://doi.org/10.1086/376718

Singer, P. (2015). *Animal Liberation*. The Bodley Head.

Tappolet, C. (2016). *Emotions, Values, and Agency*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199696512.001.0001

Tollefsen, D. P. (2015). *Groups as Agents*. Polity.

Tuomela, R. (1992). Group Beliefs. *Synthese*, *91*(3), 285–318. https://doi.org/10.1007/BF00413570

Tuomela, R. (2013). *Social Ontology: Collective Intentionality and Group Agents*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199978267.001.0001

Véliz, C. (2016). The Challenge of Determining Whether an A.I. Is Sentient. *Slate*. https://slate.com/technology/2016/04/the-challenge-of-determining-whether-an-a-i-is-sentient.html

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press

Véliz, C. (2021). Moral Zombies: Why Algorithms Are Not Moral Agents. *AI & Society*, *36*(2), 487–497. https://doi.org/10.1007/s00146-021-01189-x

Wallach, W. et C. Allen (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

Comment citer cet article: Adetou, T. E. (2025). Peut-il y avoir agentivité morale sans sentience ? *Dialogue* 64(1), 65–79. https://doi.org/10.1017/S0012217325000174

https://doi.org/10.1017/S0012217325000174 Published online by Cambridge University Press