PARAÎTRE MORAL OU L'ÊTRE : UNE CRITIQUE DES CRITÈRES D'ATTRIBUTION DE L'AGENTIVITÉ MORALE AUX INTELLIGENCES ARTIFICIELLES

MAGLO Kossi Innocent,
Université de Kara, Togo,
magloinnocent@gmail.com
BINI Essonam (MC),
Université de Kara, Togo,
e_binii@yahoo.fr

Résumé

Les systèmes d'intelligence artificielle produisent des décisions, recommandations et discours aux effets moralement significatifs. Leur aptitude à paraître moraux suffit-elle à en faire de véritables agents moraux ? À partir d'une comparaison conceptuelle et critique des conditions de l'action, de l'attribution et de la responsabilité, cet article confronte l'intentionnalisme de Kenneth Einar Himma, le fonctionnalisme de Luciano Floridi et John W. Sanders, l'approche relationnelle de Mark Coeckelbergh et la théorie de la moralité étendue de Jörg Noller. L'analyse distingue la performance morale observable, l'agentivité attribuée et la responsabilité morale forte. Elle défend un gradualisme discontinu et non substitutif : les IA actuelles ne sont ni des sujets pleinement responsables ni des instruments neutres, mais des médiations moralement actives au sein d'agencements sociotechniques. Il en résulte trois principes : la non-conversion de la performance en personnalité morale, la proportionnalité de l'attribution et la continuité de la responsabilité humaine et institutionnelle. Ces principes visent à guider la conception, l'évaluation et l'usage des IA en maintenant des interlocuteurs humains auxquels demander des comptes.

Mots-clés *: agentivité morale ; intelligence artificielle ; responsabilité ; moralité virtuelle.*

Abstract

Artificial intelligence systems produce decisions, recommendations, and discourse with morally significant effects. Is their capacity to appear moral sufficient to make them genuine moral agents? Through a conceptual and critical comparison of the conditions for action, attribution, and responsibility, this article examines Kenneth Einar Himma's intentionalism, Luciano Floridi and John W. Sanders' functionalism, Mark Coeckelbergh's relational approach, and Jörg Noller's account of extended morality. The analysis distinguishes observable moral performance, attributed agency, and robust moral responsibility. It defends a discontinuous and non-substitutive gradualism: current AI systems are neither fully responsible subjects nor neutral instruments, but morally active mediators within sociotechnical arrangements. Three principles follow: moral performance must not automatically confer moral personhood; attributed agency must be proportionate to demonstrated capacities and risks; and human and institutional responsibility must be maintained. These principles aim to guide AI design, evaluation, and use while preserving identifiable human actors from whom explanations and accountability can be required.

Keywords: *moral agency, artificial intelligence, responsibility, virtual morality*

Introduction

Une intelligence artificielle peut refuser une demande dangereuse, justifier son refus et proposer une conduite moins risquée. Une voiture automatisée peut sélectionner une trajectoire parmi plusieurs possibilités, un système d'aide à la décision médicale peut contribuer à hiérarchiser des patients et un robot d'assistance peut adapter son comportement à des signes de détresse. Dans ces situations, la machine sélectionne une action dans un espace chargé de

valeurs. La tentation est alors forte de dire qu'elle a « choisi », « compris » ou « refusé ». Cette manière de décrire son fonctionnement est pourtant équivoque : une action peut être moralement bonne sans procéder d'un agent moral, un comportement peut respecter une règle sans être accompli par devoir, et une entité peut paraître responsable sans posséder les capacités qui justifient le blâme ou la louange. Le problème consiste donc à déterminer si nous décrivons le fonctionnement de l'IA, évaluons ses performances, interprétons notre relation avec elle ou répartissons les responsabilités au sein d'un système sociotechnique.

La littérature sur les agents moraux artificiels propose des réponses divergentes. Les conceptions intentionnalistes exigent des états mentaux intentionnels, une conscience, une forme de liberté et la compréhension des raisons normatives ; Himma (2009) soutient ainsi que l'agentivité morale présuppose la conscience. Les conceptions fonctionnalistes retiennent plutôt des propriétés observables : Floridi et Sanders (2004) reconnaissent comme agents les systèmes interactifs, autonomes et adaptatifs, évaluables au regard d'un seuil moral. L'approche relationnelle de Coeckelbergh (2009, 2010) considère que la manière dont une entité apparaît et agit dans l'interaction participe à son statut moral. Enfin, les approches distribuées comprennent l'IA comme un médiateur au sein d'agencements humains, techniques et institutionnels. À partir de l'éthique

des vertus et de la cognition dite « 4E »[1], Noller (2026) l'interprète comme une stabilisation technique d'engagements humains plutôt que comme une personne vertueuse.

Ces approches ont chacune une force et un angle mort. L'intentionnalisme protège la densité du concept de responsabilité, mais risque d'imposer aux machines un modèle anthropologique difficile à vérifier même chez autrui. Le fonctionnalisme fournit des critères opérationnels, mais laisse ouverte la question du passage de la conformité à l'adhésion morale. L'approche relationnelle éclaire les pratiques d'attribution, tout en exposant la moralité à la manipulation des apparences. La distribution sociotechnique évite le mythe de l'outil neutre, mais peut dissoudre l'imputabilité dans un réseau sans répondant identifiable. Il ne s'agit donc pas de choisir une seule théorie, mais de comprendre les différentes questions auxquelles chacune tente de répondre. La question principale de cet article est alors la suivante : dans quelles conditions peut-on considérer comme un agent moral une intelligence artificielle qui semble agir moralement ? L'hypothèse défendue est qu'on peut reconnaître à l'IA une capacité d'action limitée, progressive et dépendante du contexte, sans pour autant la considérer comme un sujet moral pleinement responsable.

[1] La cognition 4E est une approche moderne des sciences cognitives qui soutient que l'esprit ne se limite pas au cerveau. Elle unit quatre dimensions complémentaires en anglais : *Embodied* (incarnée), *Embedded* (située), *Enactive* (énactive) et *Extended* (étendue).

Cette étude adopte une méthode comparative et critique fondée sur différents courants d'appréhension de l'intelligence artificielle. Le travail réunit les travaux de Himma, de Floridi et Sanders, de Coeckelbergh et de Noller, choisis parce qu'ils privilégient des critères distincts : les capacités mentales, le fonctionnement observable, la relation avec les humains et l'organisation sociotechnique de l'action. Les références à Aristote et à Kant permettent d'expliciter les principaux présupposés philosophiques de ces critères. Les travaux de Silviya Serafimova et de James H. Moor sont convoqués au fil de l'analyse pour en éclairer les limites et les conséquences.

L'analyse procède en trois étapes. Elle reconstruit d'abord les conditions que chaque approche juge nécessaires pour agir moralement ou répondre d'un acte. Les désaccords sur l'agentivité de l'IA portent notamment sur les critères utilisés, lesquels se situent à quatre niveaux différents : ce que l'IA est, ce qu'elle peut faire, la manière dont les humains la perçoivent et la responsabilité qu'on peut lui attribuer. L'analyse compare ensuite les approches à partir de quatre questions communes : que suppose-t-on de l'entité, que peut-on observer, que lui attribuent les humains et qui doit répondre de ses effets ? Elle examine enfin les conséquences de ces attributions sur la répartition des responsabilités.

Le développement suit ce fil conducteur en trois parties. La première précise les fondements philosophiques de l'action morale et examine les exigences de l'intentionnalisme. La deuxième confronte les critères fonctionnels et relationnels, en distinguant la performance

d'un système de la responsabilité qui peut lui être attribuée. La troisième articule ces résultats dans une conception graduelle et distribuée, puis en dégage trois principes pour organiser la responsabilité humaine et institutionnelle.

1. Les fondements philosophiques de la moralité artificielle

Dans la tradition aristotélicienne, l'imputation dépend du caractère volontaire de l'action et de la connaissance de ses circonstances ; l'action vertueuse engage également le choix délibéré. La vertu étant au fondement de cette morale, Aristote (1992) la définit comme :

> une sorte de milieu, puisque le milieu est le but qu'elle recherche sans cesse. […] Ainsi donc, la vertu est une habitude, un mode d'être qui dirige notre décision réfléchie, consistant dans ce milieu qui est relatif à nous, et qui est réglé par la raison comme le réglerait l'homme prudent. Elle est un milieu entre deux vices, l'un par excès, l'autre par défaut. (ARISTOTE, 1992, pp. 92-93)

La vertu n'est donc pas une succession extérieure de gestes corrects, mais une disposition stable qui fait agir de la bonne manière, pour de bonnes raisons et avec un jugement ajusté à la situation. Dans la tradition kantienne, la valeur morale ne tient pas seulement à la conformité de l'acte au devoir, mais à la maxime en vertu de laquelle le sujet agit. L'autonomie désigne la capacité de la volonté à se donner la loi, non l'indépendance technique d'un dispositif. Un système

qui exécute une règle fixée de l'extérieur peut ainsi produire une action conforme au devoir sans agir par devoir. Kant présente ainsi l'autonomie de la volonté :

> L'autonomie de la volonté est cette propriété qu'a la volonté d'être à elle-même sa loi (indépendamment de toute propriété des objets du vouloir). Le principe de l'autonomie est donc : de toujours choisir de telle sorte que les maximes de notre choix soient comprises en même temps comme lois universelles dans ce même acte de vouloir. (KANT, 2013, p. 52)

Ces différentes conceptions philosophiques expliquent la difficulté de l'expression « IA autonome ». Dans l'ingénierie, l'autonomie est ordinairement pensée sous un angle graduel. Elle mesure la capacité d'un système à fonctionner sans intervention humaine continue, à adapter ses moyens ou à gérer des événements imprévus. Dans la philosophie morale, l'autonomie est plus exigeante. Elle concerne l'origine normative de l'action et la capacité du sujet à reconnaître des raisons comme siennes. L'autonomie opérationnelle ne suffit donc pas à établir l'autonomie morale. Une voiture peut sélectionner sa trajectoire par le calcul, sans que cette opération établisse une capacité à reconnaître des fins comme siennes ou à répondre moralement de ce choix.

La responsabilité ajoute une seconde exigence. Elle va plus loin que le simple fait d'avoir causé un dommage. La responsabilité causale cherche à savoir ce qui a produit l'événement. La responsabilité liée au rôle demande qui devait surveiller ou empêcher le dommage. L'obligation de

rendre des comptes consiste à déterminer qui doit expliquer et justifier la décision. Enfin, l'imputation d'une faute morale demande dans quelle mesure l'auteur connaissait les conséquences de son action, pouvait la contrôler ou a manqué à une obligation de vigilance. Les débats sur l'IA deviennent confus lorsqu'on pense qu'une machine est coupable simplement parce qu'elle a contribué à produire un dommage. Cette distinction est au cœur du débat sur la moralité des agents artificiels.

Pour Floridi et Sanders (2004), identifier la source d'une action moralement qualifiable et déterminer si cette source mérite le blâme ou la louange sont deux opérations distinctes. Leur argument ne consiste pas seulement à relever une confusion : il conteste l'idée selon laquelle toute évaluation morale devrait se réduire à l'attribution d'une responsabilité. Une entité peut ainsi relever de l'évaluation morale sans être personnellement responsable de ses actes. Cependant, distinguer la source d'une action de son responsable ne suffit pas. Il faut déterminer si une entité peut agir moralement sans comprendre son acte ni avoir l'intention de l'accomplir. L'approche intentionnaliste soutient qu'une action pleinement morale doit procéder d'une intention consciente. Elle trouve un fondement dans la phénoménologie de Husserl, pour qui « tout état de conscience en général est, en lui-même, conscience de quelque chose » (Husserl, 1966, p. 28). Percevoir, se souvenir, imaginer ou juger, c'est toujours viser un objet selon une modalité particulière. L'acte conscient n'est donc pas un simple événement causal, mais un vécu orienté et porteur de signification. Appliquée à la morale, cette

conception distingue la conformité extérieure à une norme de l'action accomplie par un sujet capable d'en comprendre le sens et de se reconnaître comme son auteur. Un système peut-il alors être un agent moral lorsqu'il produit des comportements apparemment intentionnels sans qu'une conscience de ses actes et de leur signification morale soit établie ?

D'après la conception intentionnaliste, une action n'est pas seulement un mouvement corporel ou une sortie comportementale, mais un événement produit sous une description que l'agent peut reconnaître. Agir implique des croyances, des désirs, des intentions et une compréhension minimale de ce que l'on fait. Dans cette perspective, attribuer une moralité à une entité exige davantage que constater la régularité de ses réponses. Il faut que la règle ait pour elle une signification normative et qu'elle puisse constituer une raison d'agir. Kenneth Einar Himma (2009) défend la version la plus explicite de cette position. Selon lui, l'agentivité ordinaire présuppose que des états mentaux intentionnels causent l'action. L'agentivité morale requiert en outre une forme de choix libre ou de contrôle ainsi que la capacité de connaître et d'appliquer les prescriptions morales. Ces capacités supposent la délibération rationnelle, laquelle présuppose elle-même la conscience. La conscience n'est certes pas suffisante car un être conscient peut manquer de rationalité ou de compétence morale. Elle constitue néanmoins, dans l'argumentation de Himma (2009, pp. 20-29), une condition nécessaire de la responsabilité morale.

La question de l'agent moral artificiel dépend donc de la possibilité d'une conscience artificielle et non de la seule réussite comportementale. Carissa Véliz (2021) radicalise cette exigence en liant la compréhension morale à la sentience et à une connaissance vécue de la douleur : « pour comprendre ce que signifie infliger une douleur à quelqu'un, il est nécessaire d'avoir une connaissance vécue de la douleur » (VÉLIZ, 2021, p. 488)[2]. L'intérêt de cette position est de résister à l'inflation conceptuelle. Si l'on qualifie de « moral » un thermostat parce qu'il agit sur son environnement en respectant certaines règles, alors la notion d'agent moral devient trop large. Elle ne permet plus de distinguer une machine qui suit des consignes d'une personne capable de comprendre ses actes et d'en répondre, et qui peut donc être félicitée ou blâmée. L'intentionnalisme distingue le respect automatique d'une règle de la compréhension morale. Une machine peut éviter un dommage parce que son programme pénalise certains résultats ; un agent moral l'évite parce qu'il comprend la fragilité d'autrui. Kant précise ce point : la volonté autonome ne calcule pas seulement une conformité, elle se représente une maxime et l'éprouve au regard d'une loi qu'elle peut vouloir universelle. Une IA entraînée à produire des justifications morales ne manifeste donc pas nécessairement une volonté bonne. L'usage du mot « je » ne prouve ni une identité stable, ni la capacité de s'engager par une promesse ou de ressentir une obligation. La critique de

[2] Sauf mention contraire, les citations extraites de sources anglophones sont données dans notre traduction.

l'IA forte par Searle (1980) avait déjà signalé le risque de confondre manipulation syntaxique et saisie sémantique : maîtriser le vocabulaire moral ne suffit pas à comprendre ce qu'il exige.

Deux principales méthodes sont généralement mobilisées pour concevoir des agents moraux artificiels. La méthode descendante (top-down) part de principes ou de règles morales définis à l'avance, que le système doit appliquer aux situations rencontrées. La méthode ascendante (bottom-up), au contraire, s'appuie sur l'apprentissage à partir d'exemples, de données ou d'interactions afin que le système dégage progressivement des régularités de conduite (Wallach et Allen, 2009, p. 10). Ces deux méthodes rencontrent toutefois des limites lorsqu'il s'agit de rendre compte de la sensibilité morale. Les systèmes descendants appliquent des règles définies à l'avance, mais ces règles ne peuvent pas prévoir tous les traits singuliers susceptibles de devenir moralement importants dans une situation particulière. Les systèmes apprenants sont certes plus flexibles, puisqu'ils dégagent des régularités à partir d'exemples. Toutefois, leurs généralisations ne reposent pas sur une expérience morale partagée, façonnée par l'éducation, les relations sociales, l'empathie et la correction mutuelle. Un système peut ainsi produire une décision apparemment juste sans comprendre pourquoi un élément de la situation compte moralement, puis appliquer de manière inadéquate le même schéma à un cas nouveau. La réussite observable ne garantit donc pas une véritable compétence morale. À propos d'une compétence morale générale, Joris Graff (2024, p. 11) conclut que « ni

les agents moraux artificiels descendants ni les agents ascendants ne semblent capables d'atteindre pleinement, ou de simuler, cette sensibilité ». Il admet toutefois un rôle possible de ces systèmes dans des domaines normatifs restreints.

Jörg Noller (2026) parvient à une conclusion similaire au sujet de l'IA constitutionnelle. Même lorsqu'une intelligence artificielle apprend à respecter de manière stable certaines normes, cette régularité ne suffit pas à faire d'elle un être vertueux. Comme il le précise, « de telles régularités ne constituent ni une vertu ni un caractère moral au sens aristotélicien » (Noller, 2026, p. 12). L'intelligence artificielle peut donc se conformer aux règles et reproduire un comportement apparemment moral sans vivre les valeurs qui orientent son action ni en comprendre pleinement le sens. En effet, « les caractères moraux artificiels sont dépourvus d'affectivité, d'incarnation et de sagesse pratique (*phronèsis*) » (Noller, 2026, p. 2). Leur vertu demeure ainsi procédurale et simulée, plutôt qu'enracinée dans une expérience morale véritable.

L'approche intentionnaliste a donc un mérite important : elle permet de distinguer une machine qui applique simplement des règles d'un véritable sujet capable de comprendre et d'assumer ses actes. En exigeant la conscience, l'intention et la liberté, elle évite de considérer trop rapidement l'intelligence artificielle comme un agent moral. Cependant, ces critères soulèvent aussi plusieurs difficultés. Comment savoir ce qui se passe réellement à l'intérieur d'un autre être ? Ces exigences sont-elles appliquées de la même manière aux humains et aux machines

? Toute participation à une situation morale suppose-t-elle nécessairement une responsabilité complète ? Ces questions conduisent à examiner les principales limites de l'intentionnalisme.

La force de l'intentionnalisme ne doit pas masquer ses limites. La première est épistémologique : nous n'observons directement ni la conscience ni les intentions d'autrui, même chez les humains. Nous les inférons du comportement, des paroles, de l'histoire partagée, de la ressemblance corporelle et de pratiques communes. Himma (2009, p. 28) admet ainsi que nous pourrions devoir traiter certaines technologies sophistiquées comme des agents moraux, même sans preuve de leur conscience. Le critère interne finit donc par dépendre d'indices externes : s'il demeure absolument inaccessible, il ne permet pas de trancher ; s'il est inféré du comportement, il faut expliquer pourquoi la même inférence serait toujours illégitime pour une entité artificielle.

La deuxième limite est anthropologique. Les conditions parfois cumulées (conscience phénoménale, libre arbitre, émotions, rationalité et personnalité) forment un idéal que les humains eux-mêmes ne réalisent que partiellement. Les enfants, les personnes sous contrainte ou celles dont certaines capacités sont altérées demeurent membres de nos relations morales, même lorsque leur capacité à répondre de leurs actes est limitée. Leur statut moral ne se confond donc pas avec une pleine imputabilité. Il ne s'agit ni d'abaisser les critères jusqu'aux robots actuels ni de comparer les machines aux personnes vulnérables, mais de reconnaître que l'agentivité humaine est elle-même

soutenue par des relations, des institutions, un langage et des dispositifs matériels.

Deborah G. Johnson propose justement de considérer les systèmes informatiques comme des entités morales au sens où ils incorporent et produisent des conséquences normatives sans en faire des agents moraux. Un système de classement peut affecter des droits, redistribuer des opportunités et rendre certaines personnes plus vulnérables, sans disposer pour autant d'intentions propres. Cette position conserve la responsabilité du côté humain, mais elle oblige à reconnaître que l'artefact n'est pas moralement neutre. Ainsi, « Cela ne fait pas d'eux des agents moraux […] mais cela signifie qu'ils sont loin d'être neutres […] ce sont des entités morales » (Johnson, 2006, p. 202). Refuser une volonté à la machine ne dispense pas d'analyser la façon dont son architecture matérialise des choix de valeur.

L'intentionnalisme reste donc utile pour déterminer les conditions nécessaires afin qu'une intelligence artificielle puisse être considérée comme véritablement responsable de ses actes. Cependant, cette approche ne suffit pas à expliquer toute la place morale des IA. Elle permet de demander ce qu'une machine devrait posséder pour devenir un sujet responsable, mais elle explique moins bien comment juger son comportement, pourquoi nous avons tendance à la considérer comme un interlocuteur moral ou comment répartir les responsabilités entre les concepteurs, les utilisateurs et les institutions qui organisent le fonctionnement du système. Ces difficultés conduisent à examiner d'autres approches, notamment les conceptions fonctionnelle et relationnelle de l'agentivité morale.

Si l'intentionnalisme préserve l'exigence de compréhension et de responsabilité, peut-on cependant refuser toute agentivité morale à une intelligence artificielle au seul motif que son intériorité reste inaccessible ? Il faut donc déplacer l'analyse vers ses comportements et les relations morales qu'elle suscite.

2. De l'intériorité au comportement : les conceptions fonctionnalistes et relationnelles

Le déplacement de l'intériorité vers le comportement s'inscrit dans l'histoire de la philosophie de l'IA. Turing (1950, p. 433) proposait de remplacer la question métaphysique « les machines peuvent-elles penser ? » par le « jeu de l'imitation ». Il ne démontrait pas que toute imitation réussie est une pensée, mais recherchait un critère public d'évaluation. Appliquée à la moralité, cette démarche demande si la machine distingue les situations moralement pertinentes, justifie ses choix, révise ses réponses et se comporte de manière fiable. Une IA d'aide clinique peut, par exemple, être évaluée sur sa capacité à détecter un conflit de valeurs, à éviter les discriminations, à signaler l'incertitude, à présenter des raisons intelligibles et à solliciter une intervention humaine. Ces critères restent nécessaires même si la machine est tenue pour dépourvue de conscience ; l'incertitude métaphysique ne dispense pas de fixer des exigences aux systèmes déjà déployés.

James H. Moor (2006) distingue à cet égard plusieurs degrés : les agents à impact éthique, dont les actions produisent des conséquences morales ; les agents éthiques

implicites, conçus pour éviter certains résultats ; les agents éthiques explicites, capables de représenter et de traiter des catégories morales ; enfin les agents éthiques complets, comparables aux humains. Cette typologie évite le tout ou rien. Elle permet de dire qu'un dispositif est moralement pertinent ou fonctionnellement encadré sans affirmer qu'il est une personne.

La « moralité » mesurée dépend alors de ce que le protocole rend visible. Le comportement demeure un critère indispensable d'évaluation, mais il ne vaut pas preuve suffisante d'une intériorité ni d'une responsabilité.

Luciano Floridi et John W. Sanders (2004) proposent la formulation la plus systématique d'une agentivité artificielle indépendante des états mentaux. Leur analyse repose sur un niveau d'abstraction : une entité est décrite au moyen d'un ensemble d'observables pertinents pour une enquête. À ce niveau, un agent doit manifester trois propriétés : « L'agentivité […] dépend d'un niveau d'abstraction. Nos critères de l'agentivité sont l'interactivité, l'autonomie et l'adaptabilité […] La moralité peut être saisie comme un seuil » (Floridi et Sanders, 2004, p. 349). Il est interactif s'il répond aux transformations de son environnement et agit sur celui-ci ; autonome si son comportement peut changer sans stimulus externe immédiat ; adaptatif si les règles de transition qui gouvernent ses états peuvent elles-mêmes évoluer. Lorsque les actions d'un tel agent peuvent être évaluées à partir d'un seuil moral défini à ce niveau d'abstraction, il peut être traité comme un agent moral. Le franchissement du seuil qualifie alors une action, sans suffire à établir une responsabilité personnelle.

L'intérêt majeur de cette théorie est son non-anthropocentrisme méthodologique. Elle permet d'identifier des sources d'action dans des systèmes complexes sans supposer une copie de l'esprit humain. Dans les environnements numériques distribués, où un résultat dépend de composants logiciels, de données, d'interactions et d'apprentissages successifs, cette capacité de localisation fonctionnelle est précieuse. Luciano Floridi et John W. Sanders peuvent ainsi penser une « moralité sans esprit » et ouvrir la voie à une éthique des agents artificiels révisables, isolables ou désactivables.

Cette théorie présente néanmoins plusieurs difficultés. D'abord, le niveau d'abstraction, choisi en fonction de la finalité de l'enquête, n'est jamais neutre. Au niveau du modèle conversationnel, l'IA peut paraître autonome ; au niveau de l'infrastructure, son comportement dépend de concepteurs, d'annotateurs, de fournisseurs de données, de ressources matérielles et de politiques de déploiement. Coeckelbergh (2009, p. 188) objecte que le dépassement annoncé de l'anthropocentrisme reste construit du point de vue d'observateurs humains : « la technologie elle-même est profondément "anthropocentrique" ». Le niveau d'abstraction éclaire certains aspects du système, mais ne doit pas occulter les personnes, pratiques et choix qui rendent son action possible.

Deuxièmement, Silviya Serafimova (2020) formule cette difficulté de manière particulièrement claire : la réussite épistémique d'un agent éthique, c'est-à-dire sa capacité à calculer une solution selon une théorie kantienne,

utilitariste ou fondée sur les vertus, ne suffit pas à en faire un agent moral autonome. Il faut distinguer le calcul d'une réponse de l'estimation morale par laquelle un sujet reconnaît ce qui compte véritablement dans une situation. Cette distinction n'invalide pas la moralité fonctionnelle, mais elle en détermine les limites. Une IA peut ainsi être conçue comme un dispositif de décision soumis à des contraintes normatives, sans devenir pour autant l'auteur moral des normes qu'elle applique. Silviya Serafimova conclut ainsi que, malgré leurs différences, les trois projets d'agents moraux artificiels qu'elle examine : la machine kantienne de Powers, la réinterprétation de l'utilitarisme de l'acte par Michael Anderson et Susan Leigh Anderson, ainsi que la machine morale fondée sur l'éthique des vertus de Howard et Muntean, rencontrent une difficulté commune : passer du calcul d'une solution morale à une véritable estimation morale (Serafimova, 2020, p. 8).

Les approches relationnelles soulignent que la moralité ne s'explique pas seulement par les capacités d'une conscience isolée : elle engage aussi la relation à autrui. Dans les interactions ordinaires, nous identifions un interlocuteur à travers des signes observables : visage, voix, gestes, réponses et mémoire des échanges. Chez Levinas, toutefois, le visage ne se réduit pas à cette apparence sensible. Il redéfinit le sujet à partir de l'appel d'autrui : « L'épiphanie de l'absolument autre est visage où Autrui m'interpelle et me signifie un ordre […] C'est sa présence qui est une sommation de répondre » (Levinas, 1987, p. 53). Komi Kouvon (2019) précise cette lecture :

> En redéfinissant le sujet à partir de la responsabilité, Levinas préserve l'autonomie individuelle de toute autosuffisance et de toute attitude égocentrique et instrumentale. Mais la relation interpersonnelle telle qu'elle apparaît, dans la perspective lévinassienne, se comprend uniquement dans le sens du mouvement de l'autre vers le moi. Le visage d'autrui interpelle et convoque, le moi à la responsabilité de façon absolue. (KOUVON, 2019, p. 18)

Nous n'accédons donc pas directement à la conscience d'autrui ; nous l'appréhendons indirectement à travers sa manière d'être et d'agir. Exiger une preuve certaine de la conscience avant toute relation morale rendrait déjà problématique la reconnaissance d'autres humains comme sujets moraux. La perspective lévinassienne porte sur une autre question : l'origine éthique de la responsabilité. Pour l'intentionnalisme, le sujet est responsable parce qu'il comprend et veut son acte, puis s'en reconnaît l'auteur. Pour Levinas, la responsabilité ne procède pas seulement d'une autonomie souveraine : elle naît de l'appel d'autrui. Le moi, interpellé, est sommé de répondre. La responsabilité prend ainsi forme dans l'espace de la relation plutôt que dans une conscience isolée.

Ce déplacement permet d'interroger, par analogie, le statut moral de l'IA. Une interface dotée d'un visage, d'une voix, d'un nom, d'une mémoire et de réponses adaptées peut occuper la place d'un interlocuteur. Il faut alors examiner non seulement son éventuelle conscience, mais aussi la

manière dont elle apparaît, répond aux sollicitations et maintient une continuité comportementale. Ces médiations favorisent l'attribution d'une présence, d'intentions, voire d'une capacité à répondre de ses actes.

Coeckelbergh (2009, 2010) parle à ce propos d'agentivité morale virtuelle et de responsabilité morale virtuelle. Il ne cherche pas à prouver que les robots possèdent une conscience comparable à celle des humains, mais montre que nous attribuons l'agentivité selon la manière dont l'autre se présente, agit et répond. Une entité artificielle suffisamment cohérente peut ainsi devenir un « quasi-sujet » dans certaines interactions. Le terme « virtuel » ne désigne pas une responsabilité complète ; il décrit la manière dont nous interprétons un comportement et entrons en relation avec lui. Même sans émotions conscientes établies, un robot peut en reproduire les signes et être perçu comme un partenaire sensible.

Cette conception rappelle que la moralité d'autrui est reconnue à travers des signes concrets, explique pourquoi les utilisateurs attribuent intentions, fautes ou bienveillance aux systèmes, et attire l'attention sur leur conception. La voix, l'apparence, la rapidité de réponse, les excuses ou la mémoire des échanges influencent la relation et les attentes morales. Il faut donc demander non seulement si l'IA est morale, mais aussi comment elle nous conduit à la percevoir. Si le comportement et la reconnaissance sociale peuvent faire paraître une intelligence artificielle morale sans prouver qu'elle l'est réellement, comment articuler ces critères sans confondre performance, apparence et

responsabilité ? Cette difficulté conduit à une conception graduelle et distribuée de la moralité artificielle.

3. Vers une conception graduelle et distribuée de la moralité artificielle

La diversité de ces approches conduit à penser que l'agentivité morale peut exister à différents degrés. Cependant, cette idée doit être utilisée avec prudence. Il ne suffit pas de placer les systèmes sur une seule échelle allant du thermostat à la personne humaine. Une machine ne devient pas automatiquement responsable parce qu'elle est plus complexe ou plus autonome. En effet, les différentes capacités d'un système ne progressent pas toujours ensemble. Une IA peut, par exemple, fournir de bonnes justifications sans comprendre ce qu'est une obligation. Elle peut paraître très humaine tout en étant peu autonome sur le plan technique. Elle peut également produire des conséquences importantes sans être perçue comme l'auteur de ses actions. Il faut donc remplacer l'échelle unique par un modèle comprenant plusieurs dimensions. Cinq dimensions sont ici retenues pour former une grille qualitative d'analyse :

1. **La puissance d'intervention** : l'étendue et la gravité des conséquences que le système peut produire.
2. **L'autonomie opérationnelle** : sa capacité à agir et à s'adapter sans recevoir constamment des instructions humaines.
3. **La compétence normative fonctionnelle** : sa capacité à reconnaître certains enjeux moraux, à respecter des

règles, à expliquer ses réponses et à signaler ses incertitudes.

4. **L'agentivité attribuée** : la mesure dans laquelle les utilisateurs perçoivent le système comme un interlocuteur, un auteur ou un destinataire moral.
5. **La responsabilité morale forte** : la capacité à comprendre réellement les raisons morales, à agir en son propre nom, à maintenir une certaine continuité dans ses engagements et à se reconnaître comme l'auteur d'une faute.

Les quatre premières dimensions peuvent varier selon les systèmes et être étudiées à partir de leur fonctionnement ou de la manière dont les utilisateurs les perçoivent. La cinquième désigne les capacités qui permettent de demander à un agent de répondre personnellement de ses actes. Les performances observées ne suffisent pas, à elles seules, à établir ces capacités chez les IA actuelles. Leur dépendance à des objectifs et à des infrastructures conçus par des humains ne constitue pas non plus une réfutation suffisante de toute agentivité. Le point décisif est la différence entre produire une justification et comprendre une obligation comme sienne, ou entre recevoir une correction technique et reconnaître une faute.

Cette position peut être qualifiée de gradualisme discontinu. Avec Floridi et Sanders (2004), elle prend en compte l'interactivité, l'autonomie et l'adaptabilité sans les confondre avec la responsabilité. Avec Coeckelbergh (2009, 2010), elle reconnaît que l'attribution d'agentivité dépend des situations. Mais, en prolongeant les critiques de Himma (2009) et de Graff (2024), elle maintient qu'une amélioration

des performances ne produit pas automatiquement une volonté, une compréhension morale ou la capacité de répondre de ses actes. Une future IA responsable n'est pas déclarée impossible ; sa responsabilité ne peut simplement pas être déduite de l'accumulation de capacités techniques.

À partir de ce cadre, trois catégories pratiques peuvent être distinguées. Premièrement, les systèmes à effets moraux, qui influencent des droits, des intérêts ou des vulnérabilités sans traiter directement des normes morales. Deuxièmement, les systèmes normativement opérants, capables d'appliquer, d'apprendre ou de justifier certaines règles morales dans un domaine précis. Troisièmement, les agents moralement responsables, capables de prendre des engagements en leur propre nom et de répondre personnellement de leurs actes. Les IA actuelles peuvent relever des deux premières catégories, mais rien ne permet encore de les placer avec certitude dans la troisième. Cette classification s'inspire des quatre niveaux de James H. Moor (2006), mais les réorganise autour de la distinction entre effets moraux, traitement des normes et responsabilité personnelle. Elle constitue donc une proposition de cet article.

Un exemple hypothétique permet de préciser l'usage de cette grille. Considérons une IA qui classe des candidatures et justifie ses recommandations. Sa puissance d'intervention concerne l'accès à l'emploi ; son autonomie dépend de la place laissée à la validation humaine ; sa compétence normative s'examine à travers le respect des critères retenus et le signalement des incertitudes. L'agentivité qui lui est attribuée dépend notamment de la

confiance suscitée par ses explications. Aucune de ces observations ne suffit pourtant à établir qu'elle comprend ses obligations ou mérite personnellement le blâme. L'organisation doit donc pouvoir expliquer les critères, vérifier les recommandations et désigner les personnes habilitées à les réviser.

Behdadi et Munthe (2020) déplacent le débat vers une question pratique : dans quelles situations et jusqu'à quel point une IA peut-elle participer à des pratiques habituellement liées à la responsabilité ? Cette orientation conduit à examiner le contexte, les effets attendus et les risques. Nous en tirons ici des exigences adaptées aux usages. Dans l'éducation, traiter l'IA comme interlocuteur peut favoriser la réflexion si son caractère artificiel et ses limites sont explicites. Dans les domaines judiciaire, médical ou militaire, cette attribution ne doit jamais dispenser l'institution de justifier ses décisions. Plus une décision est grave et irréversible, plus le contrôle humain doit être effectif.

Ce cadre conduit à un premier principe : **le principe de non-conversion**. Selon ce principe, une bonne performance morale ou la reconnaissance sociale d'une IA ne doit pas être automatiquement transformée en preuve de personnalité ou de responsabilité morale. Dire qu'une IA « raisonne moralement » peut être une manière pratique de décrire son fonctionnement. Affirmer qu'elle mérite d'être blâmée exige une justification beaucoup plus forte. Ce principe empêche les humains et les institutions de transférer leurs responsabilités aux machines. Il laisse

ouverte la question distincte du statut moral de futurs systèmes dont la sensibilité pourrait être établie.

Un deuxième principe en découle : **le principe de proportionnalité attributive**. Les fonctions confiées à une IA, ses pouvoirs de décision et le vocabulaire utilisé pour la décrire doivent correspondre à ses capacités démontrées, au contexte d'utilisation et aux risques concernés. Une interface ne devrait donc pas se présenter comme consciente ou empathique si ces capacités ne sont pas établies. De même, une institution ne devrait pas utiliser une justification produite par une IA pour abandonner sa propre obligation d'expliquer et d'assumer ses décisions. Il ne s'agit pas d'interdire toute attribution d'agentivité aux machines, mais de l'encadrer avec prudence.

La dernière question n'est pas seulement métaphysique ; elle concerne aussi les institutions. Comment maintenir la responsabilité humaine lorsque les systèmes d'IA sont opaques, capables d'apprendre et intégrés dans des réseaux complexes ? Deux erreurs doivent être évitées. La première consiste à rendre la machine responsable simplement parce qu'elle a produit la décision finale. La seconde consiste à dire que l'IA n'est qu'un outil ordinaire et que nos anciennes règles de responsabilité suffisent. Or, l'apprentissage automatique et le partage des tâches entre plusieurs acteurs réduisent parfois les possibilités de prévoir et de contrôler les résultats.

La moralité étendue de Noller (2026) éclaire cette difficulté : les engagements moraux humains se stabilisent aussi dans la conception des outils, les règles d'usage et les dispositifs de gouvernance. L'analyse doit donc porter sur les

relations entre les acteurs et les systèmes. Nous en retenons que la distribution de l'action ne dispense pas de préciser les obligations de chacun. Ce qui se partage dans l'agencement sociotechnique, ce sont des contributions et des tâches ; ce partage ne transforme pas automatiquement chaque composant en sujet moral responsable.

La solution consiste à défendre un **principe de continuité de la responsabilité humaine**. Ce principe ne signifie pas qu'une seule personne doit être déclarée coupable de tous les effets imprévus d'un système. Il signifie qu'à chaque étape de son fonctionnement, des personnes ou des institutions clairement identifiées doivent être chargées d'anticiper les risques, de documenter les choix, de surveiller le système, de l'interrompre si nécessaire, d'expliquer les décisions et de réparer les dommages. L'autonomie technique peut modifier la répartition de ces obligations, mais elle ne doit jamais les faire disparaître. Comme l'écrit Helen Nissenbaum, « quel que soit le nombre d'acteurs impliqués, chacun peut être tenu de répondre également et pleinement du dommage produit » (Nissenbaum, 1996, p. 38). Cette égalité dans l'exigence de répondre ne signifie pas que tous sont également coupables, mais que la complexité du système ne doit permettre à personne d'échapper à l'obligation de répondre de sa contribution.

La vulnérabilité fournit un autre critère. Vallor et Vierkant (2024, p. 3) décrivent une « invulnérabilité structurelle » des systèmes sociotechniques aux pratiques humaines de responsabilité. Leur critique ne vise pas seulement les composants informatiques, auxquels manquent

les expériences morales et émotionnelles qui rendent sensible au reproche. Elle concerne aussi les acteurs humains que l'organisation du système éloigne des personnes touchées par ses décisions. Le gradualisme défendu est donc non substitutif : accroître l'autonomie technique ne retire pas une part équivalente de responsabilité aux humains. Il faut au contraire organiser des relations dans lesquelles les personnes affectées puissent obtenir une explication, contester une décision et demander réparation. Même lorsqu'aucun individu ne peut être blâmé pour un résultat imprévisible, l'examen institutionnel du dommage et des moyens d'y remédier demeure nécessaire.

Cette analyse reformule l'opposition entre paraître moral et être moral. Coeckelbergh (2009, 2010) reconnaît aux entités artificielles une agentivité et une responsabilité morales virtuelles. Une IA peut apparaître comme un interlocuteur capable de formuler des raisons, de respecter des normes ou de répondre de ses actes, sans que cela établisse une conscience morale comparable à celle de l'être humain. Denis Berthier (2004, p. 73) définit le virtuel comme « ce qui, sans être réel, a, avec force, les qualités de [la chose réelle] ». Une moralité virtuelle n'est donc ni pleine ni dépourvue d'effets : l'IA peut reconnaître une situation sensible, appliquer une règle, justifier une réponse ou adapter son comportement, sans devenir pour autant une personne moralement responsable.

La question essentielle n'est donc pas seulement de savoir si les IA sont morales, mais quelle place morale leur reconnaître et quelles responsabilités humaines doivent accompagner leur développement. La prudence

aristotélicienne invite à éviter deux excès : les humaniser et les réduire à des outils neutres. La position défendue reconnaît une moralité virtuelle, fonctionnelle et évolutive sans la confondre avec la responsabilité pleine d'une personne. Plus les IA disposent de capacités d'action et de décision, plus la responsabilité humaine et institutionnelle doit être explicitement organisée.

Conclusion

L'analyse permet de répondre à la question initiale : paraître moral ne suffit pas à être un sujet moral pleinement responsable. Les IA peuvent produire des effets moralement significatifs, appliquer des règles et être perçues comme des interlocuteurs. Ces capacités justifient l'examen de leur rôle dans l'action, mais elles ne démontrent ni une compréhension des obligations ni la capacité de répondre personnellement d'une faute. La distinction entre performance, attribution et responsabilité doit donc guider l'évaluation de ces systèmes.

La comparaison montre l'apport et la limite de chaque approche. L'intentionnalisme précise les exigences de la responsabilité, mais rencontre la difficulté d'accéder à l'intériorité d'autrui. Le fonctionnalisme rend les conduites évaluables, sans établir à lui seul une adhésion morale. L'approche relationnelle explique la place des apparences et des interactions, sans garantir une responsabilité pleine. La moralité étendue invite enfin à examiner l'organisation des relations entre humains, outils et institutions. Ces perspectives éclairent des dimensions distinctes du

problème ; leur articulation ne supprime pas leurs désaccords.

Le gradualisme discontinu et non substitutif proposé dans cet article tient ensemble ces résultats. Les capacités d'intervention, d'adaptation et de traitement des normes peuvent varier en degré, tandis que la responsabilité personnelle exige une justification distincte. Reconnaître le rôle moral des IA ne revient donc pas à leur transférer les obligations des humains. Les trois principes de non-conversion, de proportionnalité de l'attribution et de continuité de la responsabilité humaine et institutionnelle donnent un contenu pratique à cette distinction. L'utilité de cette étude est d'offrir des repères aux concepteurs, aux utilisateurs et aux institutions. Les concepteurs doivent adapter les fonctions et la présentation du système à ce qu'il peut réellement faire. Pour les organisations, il s'agit de préciser qui choisit les critères, contrôle les résultats, peut interrompre l'usage et répond aux contestations. Pour l'enseignement et la formation, cette distinction aide à évaluer une justification produite par l'IA sans lui attribuer d'emblée une compréhension morale. Elle permet enfin aux personnes concernées par une décision de savoir à quel interlocuteur humain ou institutionnel demander des comptes. Cette proposition reste une grille philosophique. Elle ne fournit ni un test de conscience artificielle ni une mesure universelle de l'agentivité. Sa pertinence pratique devra être examinée dans des situations précises, en tenant compte des capacités des systèmes, des moyens réels de contrôle et de l'expérience des personnes affectées. La

question de l'éventuelle reconnaissance de futurs agents artificiels demeure toutefois ouverte.

L'enjeu est ainsi de rendre les usages de l'IA discutables et les responsabilités identifiables. Plus ces systèmes participent aux décisions, plus les conditions permettant aux humains de les expliquer, de les contester et d'en assumer les conséquences doivent être organisées. La reconnaissance d'une moralité fonctionnelle ou virtuelle prend alors son sens dans la préservation d'une responsabilité humaine effective.

Références bibliographiques

ARISTOTE, 1992. *Éthique à Nicomaque*, trad. J. Barthélemy-Saint-Hilaire, révision par A. Gomez-Muller, Librairie générale française, Paris.

BEHDADI Dorna et MUNTHE Christian, 2020. « A normative approach to artificial moral agency », *Minds and Machines*, vol. 30, n° 2, pp. 195-218, https://doi.org/10.1007/s11023-020-09525-8, consulté le 14 septembre 2026.

BERTHIER Denis, 2004. *Méditations sur le réel et le virtuel*, L'Harmattan, Paris.

COECKELBERGH Mark, 2009. « Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents », *AI & Society*, vol. 24, n° 2, pp. 181-189, https://doi.org/10.1007/s00146-009-0208-3, consulté le 14 septembre 2026.

COECKELBERGH Mark, 2010. « Moral appearances: Emotions, robots, and human morality », *Ethics and Information Technology*, vol. 12, n° 3, pp. 235-241, https://doi.org/10.1007/s10676-010-9221-y, consulté le 14 septembre 2026.

FLORIDI Luciano et SANDERS John W., 2004. « On the morality of artificial agents », *Minds and Machines*, vol. 14, n° 3, pp. 349-379, https://doi.org/10.1023/B:MIND.0000035461.63578.9d, consulté le 14 septembre 2026.

GRAFF Joris, 2024. « Moral sensitivity and the limits of artificial moral agents », *Ethics and Information Technology*, vol. 26, article 13, https://doi.org/10.1007/s10676-024-09755-9, consulté le 14 septembre 2026.

HIMMA Kenneth Einar, 2009. « Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? », *Ethics and Information Technology*, vol. 11, n° 1, pp. 19-29, https://doi.org/10.1007/s10676-008-9167-5, consulté le 14 septembre 2026.

HUSSERL Edmund, 1966. *Méditations cartésiennes. Introduction à la phénoménologie*, trad. Gabrielle Peiffer et Emmanuel Levinas, J. Vrin, Paris.

JOHNSON Deborah G., 2006. « Computer systems: Moral entities but not moral agents », *Ethics and Information Technology*, vol. 8, n° 4, pp. 195-204, https://doi.org/10.1007/s10676-006-9111-5, consulté le 14 septembre 2026.

KANT Emmanuel, 2013. *Fondements de la métaphysique des mœurs*, trad. Victor Delbos, Les Échos du Maquis, édition électronique, https://philosophie.cegeptr.qc.ca/wp-content/documents/Fondements-de-la-M%C3%A9taphysique-des-moeurs.pdf, consulté le 14 septembre 2026.

KOUVON Komi, 2019. *La responsabilité éthique dans les sociétés postcommunicationnelles*, L'Harmattan, Paris.

LEVINAS Emmanuel, 1987. *Humanisme de l'autre homme*, Librairie générale française, Paris.

MOOR James H., 2006. « The nature, importance, and difficulty of machine ethics », *IEEE Intelligent Systems*, vol. 21, n° 4, pp. 18-21, https://doi.org/10.1109/MIS.2006.80, consulté le 14 septembre 2026.

NISSENBAUM Helen, 1996. « Accountability in a computerized society », *Science and Engineering Ethics*, vol. 2, n° 1, pp. 25-42, https://doi.org/10.1007/BF02639315, consulté le 14 septembre 2026.

NOLLER Jörg, 2026. « Artificial moral characters: Constitutional AI and the challenge of alignment », *AI and Ethics*, vol. 6, article 196, https://doi.org/10.1007/s43681-026-01033-2, consulté le 14 septembre 2026.

SEARLE John R., 1980. « Minds, brains, and programs », *Behavioral and Brain Sciences*, vol. 3, n° 3, pp. 417-424, https://doi.org/10.1017/S0140525X00005756, consulté le 14 septembre 2026.

SERAFIMOVA Silviya, 2020. « Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement », *Humanities and Social Sciences Communications*, vol. 7, article 119, https://doi.org/10.1057/s41599-020-00614-8, consulté le 14 septembre 2026.

TURING Alan M., 1950. « Computing machinery and intelligence », *Mind*, vol. 59, n° 236, pp. 433-460, https://doi.org/10.1093/mind/LIX.236.433, consulté le 14 septembre 2026.

VALLOR Shannon et VIERKANT Tillmann, 2024. « Find the gap: AI, responsible agency and vulnerability », *Minds and Machines*, vol. 34, article 20, https://doi.org/10.1007/s11023-024-09674-0, consulté le 14 septembre 2026.

VÉLIZ Carissa, 2021. « Moral zombies: Why algorithms are not moral agents », *AI & Society*, vol. 36, n° 2, pp. 487-497, https://doi.org/10.1007/s00146-021-01189-x, consulté le 14 septembre 2026.

WALLACH Wendell et ALLEN Colin, 2009. *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, New York.