Réflexion sur la responsabilité de l'IA à partir de la théorie de la responsabilité de Hans Jonas

Basile NGONO
Maître de Conferences
École Nationale Supérieure Polytechnique, Yaoundé, Cameroun
amingono@gmail.com
Roger TAMBANGA
Doctorant au Laboratoire de Philosophie (LAPHI)
Université Joseph KI-ZERBO, Ouagadougou, Burkina-Faso
rogertambanga@gmail.com

Résumé :

Les progrès enregistrés ces derniers temps dans le domaine de l'intelligence artificielle suscitent de nombreux débats éthiques. L'avènement des robots, des véhicules autonomes et la mise en perspective des IA fortes nous conduit aujourd'hui à nous demander si l'humanité doit déléguer une partie de sa responsabilité ou se la décharger complètement au profit de l'IA. Considérée comme un objet de responsabilité pour certains et un sujet de responsabilité pour d'autres, l'IA nous invite à questionner à nouveau le concept de la responsabilité. La réflexion qui se déploie ici entend penser le sujet de la responsabilité de l'IA à la lumière de l'éthique de la responsabilité de Jonas, étant entendu que Jonas est le philosophe contemporain qui a théorisé de façon systématique sur le concept de responsabilité en lien avec les progrès technologiques. Eu égard des fondements ontologico-métaphysique et théologico-philosophique, et vu le telos de la responsabilité tels que pensés par Jonas, il apparait saugrenu de parler de la responsabilité de l'IA. La responsabilité, dans sa version jonassienne, serait le Rubicon que l'IA ne pourrait franchir. Il s'agit, bien au contraire, avec le succès de l'IA, d'accroître la responsabilité humaine.

Mots clés : Intelligence Artificielle, métaphysique, ontologie, responsabilité, théologie.

Abstract:

Recent progress in the field of artificial intelligence has given rise to numerous ethical debates. The advent of robots, autonomous vehicles and the prospect of strong AI leads us today to ask ourselves whether humanity should delegate part of its responsibility or offload it completely for the benefit of AI. Considered as an object of responsibility for some and a subject of responsibility for others, AI invites us to question the concept of responsibility again. The reflection that unfolds here intends to think about the subject of the responsibility of AI in the light of Jonas's ethics of

responsibility, as Jonas is the contemporary philosopher who has theorized systematically on the concept of responsibility in link with technological progress. Given the ontological-metaphysical and theological-philosophical foundations, and given the telos of responsibility as thought by Jonas, it appears preposterous to speak of the responsibility of AI. Responsibility, in its Jonas version, would be the Rubicon that AI could not cross. On the contrary, with the success of AI, it is about increasing human responsibility.

Keywords: Artificial Intelligence, metaphysics, ontology, responsibility, theology.

Introduction

Le monde est en passe de vivre une nouvelle révolution qu'on pourrait appeler la révolution de l'intelligence artificielle. Si les premiers systèmes intelligents remontent aux années cinquante (en faisant référence au Test de Turing), il faut reconnaître qu'on assiste à un réel progrès dans le secteur de l'IA à partir de ces deux dernières décennies. Un progrès qui sera redevable aux progrès des neurosciences, à la performance des ordinateurs, aux *big data*. Les domaines d'applications de l'IA se sont également élargis. Son *telos* est en train de se transformer radicalement, car de simples assistants, les IA sont en passe de jouir du statut d'agents autonomes". L'agencité de l'IA suscite aujourd'hui d'inquiétudes et de nombreuses interrogations éthiques. Avec l'IA forte, le champ de la responsabilité humaine se rétrécit et paradoxalement s'élargit. Si le débat sur la responsabilité de l'IA se pose, cela fait suite à la mise en circulation des véhicules intelligents sinon "autonomes", à l'introduction des robots pour personnes âgées, à la possible mise en service des armes intelligentes ou des robots-tueurs/militaires, des robots juges, soigneurs qui pourront prendre des décisions de manière autonome. Mais s'agit-il d'une autonomie faible, ou forte dans le sens du sujet autonome kantien ? En cas d'échec d'un système d'intelligence artificielle autonome, à qui responsabilise : l'entreprise, l'utilisateur, l'IA, ou la société dans son ensemble ? Et jusqu'où peut aller cette responsabilité, si c'est l'entreprise, l'utilisateur, l'IA ou la société qui devrait l'assumer ? Voilà déclinées quelques questions en lien avec la responsabilité de l'IA dans ses applications actuelles et envisagées dans un futur proche au lointain. On pourrait bien constater que les

débats en cours se cristallisent sur l'agent à qui imputer la responsabilité. Une posture qui conduit à s'intéresser de moins en moins aux fondements et aux conditions de possibilité de la responsabilité. Hans Jonas fut le premier à poser la question de la responsabilité en lien étroit avec les progrès technologiques. Un nouveau type de responsabilité à l'âge technologique qu'il s'est évertué à penser les fondements et la portée. Dans le sillage du débat éthique sur la responsabilité de l'IA, on est amené à se demander si le principe de responsabilité de Jonas ne pourrait pas éclairer un tel débat. Par-delà les questions sus-soulevées, la question centrale qui oriente la réflexion en cours est la suivante : serait-il approprié de parler de la responsabilité de l'IA à partir du principe de responsabilité de Hans Jonas ? L'hypothèse qui sous-tend notre réflexion est que l'IA ne peut assumer des responsabilités sur des sujets humains. La vérification d'une telle hypothèse, nous conduira à adopter une approche philosophique fondée sur une méthode herméneutique analytique et critique qui s'appuie sur la littérature de l'IA et sur les textes de Jonas. L'intérêt de la réflexion vise à montrer que le succès de l'IA accroît la responsabilité humaine au lieu de l'amenuiser. Pour ce faire, notre réflexion se construira autour des points suivants : il s'agira de s'employer, dans une première étape, à un exercice de clarification conceptuelle, dans un deuxième moment à présenter les fondements du principe de responsabilité de Jonas et en dernière instance à procéder à une discussion sur la capacité de responsabilité de l'IA.

1. IA, sens et domaines d'application

Qu'est-ce que l'intelligence artificielle ? Voilà une question qui pourra, à première vue, recevoir une réponse univoque et simplifiée. Mais au regard des réponses tant plurielles que différentes rapportant à la question, on se rend vite compte que le sens de la notion de va pas de soi. Il faut commencer par observer que le concept d'IA est un concept qui revêt plusieurs sens, car s'appuyant sur des théories et orientations différentes. Procéder à un détour d'analyse critique des différents sens que revêt le mot s'impose ici.

La locution d'intelligence artificielle revêt plusieurs sens. Elle est employée pour faire référence à une science qui s'évertue à concevoir

et programmer des systèmes artificiels dotés d'intelligence. Des fonctions intelligentes conçues par des experts qui seront ensuite implanter dans des systèmes (logiciels, programmes…) ou objets (ordinateurs, robots…). Comme le fait observer Aurélien Vannieuwenhuyze (2019, p. 27) « Les robots, les voitures autonomes ne sont pas ce que l'on peut appeler des intelligences artificielles, ce sont des machines utilisant de cette intelligence ». La définition que propose Vannieuwenhuyze fait sans doute allusion aux systèmes experts. Ici, la démarche technique consiste à formaliser des comportements humains de sorte qu'ils puissent être dupliqués dans une machine. Se fondant sur des principes logico-mathématiques, les systèmes experts requièrent que soit transcrit en langage et interférence symboliques le raisonnement humain. La raison calculatrice est prise ici pour ce qui est de tout intelligent. Écartant la vie subjective, l'IA n'est plus qu'un agent rationnel, s'efforçant d'atteindre un objectif de façon optimale. Comme l'écrivent justement R. Stuart et N. Peter (2021, p. 16), « Un agent rationnel est un agent qui agit de manière à atteindre le meilleur résultat ou, dans un environnement incertain, le meilleur résultat espéré ». Mais une intelligence qui prétend mettre entre parenthèses la vie sentimentale, émotive, peut-elle encore mériter l'appellation d'intelligence, quoique artificielle ? L'homme est autant rationnel que raisonnable. Ce qui veut dire que l'IA dans les systèmes experts n'est pas un sujet moral. De tels systèmes rencontreront des difficultés quasi insurmontables, car il n'est pas possible de traduire la richesse des comportements humains, surtout ceux non rationnels (sentiments, émotions, créativité, sens moral…) par des symboles formels qui puissent être implémentés dans des machines. Une incapacité qui pourrait conduire à la conclusion que l'IA symbolique n'est pas une intelligence *stricto sensu*. Alors d'où tient le succès de l'IA au point de susciter des inquiétudes éthiques ?

Un autre sens qui revient est qu'on entend par IA, tout système artificiel permettant de simuler en partie ou en tout l'intelligence humaine. « La définition largement acceptée de l'intelligence artificielle est celle d'un programme informatique étant à même de reproduire des fonctions cognitives associées à l'esprit humain : apprendre, interagir, déduire, raisonner… » (J.-P. Desbiolles & G. Colombet, 2023, p. 30). Puisqu'il s'agit de simuler les fonctions

cognitives de l'homme, les concepteurs et développeurs des IA vont s'efforcer de construire des réseaux neuronaux à même d'imiter parfaitement le cerveau humain. L'IA est dite faible lorsqu'elle imite une parcelle d'intelligence humaine, et forte quand elle peut faire montre d'une intelligence générale et de conscience humaine. L'avantage que le système neuronal a sur l'IA symbolique est que le système évolue dans le champ de l'imprévisible, de l'indéterminé. Par le *deep learning* ou par l'apprentissage par renforcement, le système neuronal, à plusieurs couches, parviennent à des résultats extraordinaires.

Il faut tout de suite observer que la simulation dont il est question est à l'heure actuelle partielle. Pour l'heure, nous avons affaire dans la pratique à des IA spécifiques. Le réseau neuronal, quand bien même il s'inspire du cerveau biologique humain et éclairer par des neurosciences, n'est pas encore une IA générale. Les limites seraient moins techniques que théoriques. Tout porte à penser ici que le cerveau serait le siège absolu de l'intelligence. Ce qui conduit à sous-estimer l'ensemble du corps biologique de l'homme dans le discernement de soi et du monde. Par-delà le dualisme et le monisme réducteur, l'intelligence humaine ne peut être domiciliée dans une seule région particulière de l'homme. On est même tenté de dire que toutes les parties de l'organisme sont coproductrices de l'intelligence humaine. La critique de Dreyfus (1984) dirigée contre l'IA s'inscrit dans cette approche de l'intelligence humaine. Ce qui n'est pas vu comme une intelligence chez les humains est considéré en revanche comme une intelligence extraordinaire quand il est reproduit par une machine. Il s'agit de la perception, du langage, de la mobilité etc. Cette amplification des résultats des IA spécifiques conduit certains laudateurs à les envisager autrement.

Enfin, on parle de l'IA pour faire référence à la superintelligence. La superintelligence évoque le titre d'un ouvrage de Nick Bostrom. Dans cette approche, on part du présupposé que l'intelligence humaine est limitée et peut être améliorée, sinon être transcendée. La superintelligence s'inscrit dans le projet de la singularité dont l'une des figures de proue est Ray Kurzweil. Pour cet ingénieur, adviendront dans un futur proche des machines super-intelligentes capables de s'auto-reproduire. Ray Kurzweil prédit, en 2045, l'avènement « d'une intelligence non-biologique un milliard de

fois plus performant que l'esprit humain. Cet événement charnière de l'histoire de l'humanité, il l'appelle "singularité". (…) La singularité devrait, selon lui, survenir à partir du moment où la technologie deviendra si complexe, si autonome, si intelligente, qu'elle prendra en charge sa propre évolution dans un mouvement de croissance ininterrompu et exponentiel » (D. Neerdael, 2018, p. 439). Une telle conception de l'IA, même si elle se nourrit actuellement de la convergence des NBIC (Nanotechnologie, Biotechnologie, Science Informatique et Science Cognitiviste), relève de la science-fiction.

Il convient de noter que les définitions retenues ici n'épuisent pas le champ sémantique du mot. La réflexion en cours privilégiera les définitions qui présupposent l'autonomie des systèmes d'intelligence artificielle (SIA).

Au vu de leurs succès réels et présumés, les IA s'invitent dans tous les secteurs d'activité de l'homme. Et ne pouvant pas, dans le cadre de ce présent travail, nous intéresser à tous ses domaines d'application, nous focaliserons notre attention sur ses champs d'application tels que le domaine militaire, le domaine de la santé et enfin celui du transport. Ce sont des domaines où confier une décision à teneur morale à une IA dans son interaction avec les humains est plus que problématique. Dans le domaine militaire, on envisage l'usage des armes intelligentes ou des robots-tueurs. Si demain ces robots venaient à prendre part aux conflits armés ou non-armés, ils décideront selon leur dessein propre de qui vivra ou non. Le droit à la vie des humains dépendra peut-être de la "volonté" de ces robots. Dans le transport, il s'agit déjà de la mise en circulation des véhicules autonomes. Des véhicules dont la "responsabilité" s'étend de la sécurité du "propriétaire" ou des passagers à celle des autres usagers de la route. Et enfin de "robots médecins" autonomes décideront qui diagnostiquer et qui peut bénéficier des soins préventifs ou curatifs. Dans tous les cas, la responsabilité des IA est posée quand on envisage de les rendre comme des agents autonomes dotés d'une volonté. Les hommes doivent-ils déléguer leur responsabilité aux SIA ? Avant tout, quelles sont les conditions de possibilité de l'exercice de la responsabilité ?

2. Fondements et conditions de possibilité de la responsabilité chez Hans Jonas

Le développement de l'IA appelle-t-il la responsabilité humaine ou la responsabilité des machines ? Avant de s'employer à répondre à cette question nodale de notre réflexion, tâchons-nous à dégager les fondements de la responsabilité chez Jonas.

Commençons par remarquer que la responsabilité chez Jonas ne repose pas sur un fondement juridique. La particularité de la responsabilité juridique est qu'elle repose sur une faute déjà commise et sanctionnée par les lois de la cité. L'imputabilité d'un acte posé après coup constitue le sens de la responsabilité juridique. Alors que la responsabilité morale jonassienne se définit « non comme l'imputation à un individu des actes qu'il a accompli, mais comme une exigence de limitation d'un pouvoir » (J-Y. Goffi, 1993, p. 208). À partir de cette remarque préliminaire on peut déjà penser que la responsabilité de l'IA serait probablement une responsabilité pour faute commise et non une responsabilité fondée sur un pouvoir à en commettre une. Nous reviendrons sur la distinction que Jonas instaure entre la responsabilité morale et la responsabilité juridique.

L'originalité de la responsabilité jonassienne tient du fait qu'elle se fonde sur des soubassements théologico-philosophiques[1]. S'inspirant d'un mythe de la création que lui-même a construit, Jonas en vient à indiquer que l'homme est cet être à qui, il incombe la « responsabilité envers le monde réel et les désastres que sa frivolité ou l'ignorance des conséquences de ses actes peuvent causer ; ensuite sa responsabilité plus essentielle à l'égard du devenir de la transcendance ; il y a enfin sa responsabilité à l'effet en retour qu'une transcendance menacée peut exercer sur les hommes » (M. Weyembergh, 1993, p. 191).Nous ne pourrons pas, dans les limites de ce travail, revenir dans les détails sur le mythe de la création cosmique que Jonas expose dans *Le Concept de Dieu après Auschwitz.*

[1] Il faut comprendre par fondement théologique non pas un fondement qui s'appuie sur des considérations religieuses consacrées, mais par fondement théologie une construction philosophique et rationnelle mettant en scène l'idée de Dieu dans la saisie des phénomènes aussi bien cosmiques qu'humains. Jonas écarte l'idée de fonder sa théorie de la responsabilité sur une foi religieuse.

Trois traits caractéristiques de l'idée de Dieu retiennent notre attention dans ce mythe. Il s'agit d'un Dieu en devenir, en souffrance et d'un Dieu impuissant. Sans ces traits caractéristiques, le monde devant nous, les visages humains que nous connaissons et bien d'autres *étants,* pour emprunter une terminologie heideggérienne, *ne* sauraient advenir. L'existence du monde et des êtres qui le composent est la conséquence du renoncement à l'essence de Dieu, à son *existentialité-essentialité*. « Dieu pour que le monde soit et qu'il existe de par lui-même, a renoncé à son Etre propre ; il s'est dépouillé de sa divinité, afin d'obtenir celle-ci, en retour, de l'odyssée des temps, donc chargée de la récolte fortuite d'une imprévisible expérience temporelle, lui-même, Dieu, étant alors transfiguré, ou peut-être aussi défiguré par elle. » (H. Jonas, 1994, p. 15) Un tel dépouillement rend Dieu vulnérable dans cette odyssée ; une vulnérabilité accompagnée de son impuissance. « Au concept d'un dieu souffrant et d'un dieu en devenir se trouve étroitement lié le concept d'un dieu *soucieux* – non pas éloigné, détaché, en lui-même enfermé, mais au contraire impliqué dans ce dont il a le souci » (H. Jonas, 1994, p. 26). Et Jonas (1994, pp. 26-27) d'ajouter, « ce Dieu soucieux n'est pas un magicien qui, par le seul acte de son souci, provoquerait simultanément la réalisation du but dont il a le souci : au contraire, il a laissé à d'autres acteurs quelque chose à faire, de sorte que son souci dépend d'eux. C'est donc aussi un dieu en péril, un dieu qui encourt un risque propre. (…) ce Dieu-là n'est pas un dieu tout-puissant !».

L'apparition de la vie subjective, consciente et spirituelle donne l'occasion à la divinité de se sauver du risque qu'il a décidé d'encourir. En l'homme, Dieu se redécouvre, retrouve partiellement sa puissance et se remet de son ignorance destinale. Cette puissance partielle retrouvée le maintient cependant dans l'angoisse existentielle et essentiale, car la liberté humaine peut stopper à tout moment cette odyssée divine. « La transcendance s'est éveillée à elle-même avec l'apparition de l'homme, et elle accompagne désormais les actions de ce dernier en retenant son souffle, avec l'espoir du demandeur, dans la joie et dans la tristesse, dans la satisfaction et dans la désillusion, se rendant, comme j'aimerais le croire, sensible à lui, sans pourtant intervenir dans la dynamique du théâtre du monde (…). » (H. Jonas, 1994, pp. 20-21)

L'implication éthique que nous tirons de l'aventure cosmique, ou ce qui revient au même de l'aventure divine, est que l'homme devient cet être à qui échoit la responsabilité de prendre soin de la créature et d'un Dieu s'efforçant de se réconcilier avec lui-même. En l'homme est confié une très grande responsabilité, puisque « Dieu, après s'être entièrement donné dans le monde en devenir, n'a plus rien à offrir : c'est maintenant à l'homme de lui donner. Et il peut le faire en veillant à ce que, dans les cheminements de sa vie, n'arrive pas, ou n'arrive pas trop souvent, et pas à cause de lui, l'homme, que Dieu puisse regretter d'avoir laissé devenir le monde » (H. Jonas, 1994, pp. 38-39).

À partir de ce mythe, on se rend compte que la responsabilité de l'homme se justifie par la puissance dont il dispose et de l'impuissance divine. Il ne s'est pas agi ici d'évoquer des principes religieux, en particulier d'une religion transcendante qui consacre l'homme comme un intendant du monde créé, mais d'une reconstruction de l'idée de Dieu pour fonder la responsabilité humaine. Une première condition de la responsabilité est donc le dépouillement de Dieu de son essence. L'arrière-plan théologico-philosophique de la responsabilité est prolongé par l'approche ontologico-métaphysique. Si la responsabilité humaine découle de l'impuissance divine, faut-il, *mutatis mutandis*, penser la responsabilité de l'IA comme résultant de l'impuissance humaine ? Avant de s'atteler à apporter une réponse à cette interrogation, examinons le second fondement de la responsabilité jonassienne.

Dans son ouvrage majeur, *Le Principe Responsabilité,* Jonas s'emploie à fonder ontologiquement et métaphysiquement sa théorie de la responsabilité. Ce qui a été esquissé comme fondements de la responsabilité dans *Le Concept de Dieu après Auschwitz,* dans le *Phénomène de la vie* et dans *Puissance et Impuissance de la subjectivité* trouve son achèvement dans *Le Principe Responsabilité*. Ce qu'il faut déjà observer est que le fondement théologico-cosmique est solidaire du fondement ontologico-métaphysique. La théorie de la responsabilité s'ouvre par ces mots : « Fonder le "Bien" ou la "Valeur" dans l'être, cela veut dire enjamber le prétendu gouffre entre l'être et le devoir » (H. Jonas, 1990, p. 157). Pour Jonas, le sens de responsabilité est domicilié en l'homme et son appel est adressé par des êtres vulnérables se trouvant sous le pouvoir de son action.

Autrement dit, la responsabilité a un fondement ontique et ontologique. Les concepts métaphysiques que Jonas mobilise dans cette fondation de la responsabilité sont essentiellement les concepts de fins, de valeurs, de vulnérabilité et de liberté. Les êtres vivants qui peuvent être impactés par le faire de l'homme, pour parler comme Jonas, sont des êtres poursuivant des fins. Poursuivre un but donné témoigne de la valeur de la chose poursuivie. « En entretenant des fins ou en ayant des buts, comme nous le supposerons maintenant, la nature pose également des valeurs ; car devant une fin donnée de quelque manière que ce soit et recherchée de *facto*, son obtention dans chaque cas devient un bien et son empêchement un mal, et avec cette différence commence l'imputabilité de la valeur » (H. Jonas, 1990, p. 158). Le devoir-être de l'Etre s'explique par la valeur de l'Etre sur le non-Etre. Après avoir dégagé objectivement des valeurs dans la nature à partir du concept de fin, Jonas estime qu'il appartient à l'homme de protéger la valeur de l'Etre, et ce d'autant plus qu'il dispose d'un pouvoir capable de mettre en péril cet Etre-vulnérable. Autrement dit, l'homme étant au sommet du règne des fins et connaissant la valeur des fins et ayant intérêt que se perpétue la possibilité de poser et de poursuivre des fins, devient cet être à qui revient la responsabilité de garantir la possibilité de responsabilité au monde. Sans créer une rupture ontologique entre l'homme et le monde de la vie, Jonas reconnait tout de même que l'homme se trouve au sommet de la vie téléologique.

La subjectivité consciente de l'homme, la capacité de transcendance de ce dernier, la prédisposition ontique de l'homme d'entendre l'appel à la responsabilité que lui adresse le monde de la vie et la liberté humaine, constituent les conditions de possibilité de la responsabilité chez Hans Jonas. Une responsabilité qui prend en charge l'avenir de l'humanité et de la vie en générale. Et surtout une responsabilité qui a pour « fin le maintien de l'exercice d'une responsabilité dans l'avenir, dans la mesure où seule cette responsabilité est le gage d'une autonomie souveraine » (E. Pommier, 2013, p. 139).

En résumé, l'impuissance divine et la valeur de l'être d'une part, et d'autre part la transcendance, la liberté et la subjectivité consciente de l'homme, fondent la responsabilité jonassienne. A partir de telles conditions de possibilité de la responsabilité, peut-on à

proprement parler de la responsabilité de l'IA, ou convient-il de parler d'un développement responsable de l'IA ?

3. Responsabilité de l'IA en question

Le contexte ayant conduit Hans Jonas a élaboré sa théorie de la responsabilité est bien évidemment celui marqué par des progrès technoscientifiques. Avec les nouveaux pouvoirs que confère la technique à l'homme aujourd'hui, il y a nécessité de repenser la responsabilité. De prime apport, la réflexion de Hans Jonas ne s'adresse par en particulier aux intelligences artificielles. Se prononçant sur les problématiques en lien avec les systèmes d'IA de son temps, le philosophe fera remarquer : « Cela ne peut donc nullement intéresser particulièrement un philosophe ». Il laisse entendre ici que l'IA ne peut susciter un intérêt philosophique. Alors notre projet d'examiner la pertinence de la responsabilité de l'IA à partir du principe de la responsabilité ne devient-il pas une entreprise risquée, sinon vouée à l'échec ? Le désintérêt du philosophe au sujet des IA s'expliquerait par le fait que le niveau de développement de l'IA à l'époque était à son stade de balbutiement. Les années 80 étaient témoins de l'avènement des systèmes experts (IA symbolique) et de l'hiver de l'IA. Cependant les projections envisageaient déjà l'avènement des IA autonomes.

Même si Hans Jonas ne s'est pas consacré à l'élaboration d'une réflexion systématique sur l'IA, il ne manque pas dans sa réflexion des pistes pour penser les nouveaux défis éthiques qu'engendrent les IA. Ayant mené une réflexion globale sur les progrès technologiques, l'éthique jonassienne de la responsabilité nous offre des ressources intellectuelles pour soit poser les nouvelles questions éthiques en lien avec les IA soit répondre aux inquiétudes éthiques que ces dernières suscitent. Aux questions : « Ne peut-on pas dire que lorsqu'on fait appel aux ordinateurs ou aux réseaux de neurones pour arrêter une décision, ceux-ci assument des fonctions qui sont du moins équivalentes à une participation à la responsabilité ? Les hommes ne sont-ils pas dans ce cas-là de *facto* conduits à déléguer une responsabilité inconfortable aux machines ? » (H. Jonas, 2017 p. 103), la réponse du philosophe était sans équivoque. Pour lui :

> Les décisions sont prises par des sujets qui assument également la responsabilité. Une machine ne pourra jamais la leur ôter. (…) si l'introduction de systèmes informatiques équivalents devait avoir pour conséquence que l'importance capitale de l'individu dans sa singularité soit minée au profit d'une machinerie sociale qui, au mieux, travaillerait sans heurt, cela serait grave : la perte du respect à l'égard de la subjectivité constituerait effectivement un grand danger pour l'humanité » (H. Jonas, 2017 pp. 103-104).

Il ressort de ces mots de Jonas que l'on ne peut parler de responsabilité que lorsqu'on a affaire à un sujet. Un sujet dont les implications philosophiques sont entre autres : une conscience consciente d'elle-même, une volonté, une transcendance subjective, une capacité d'émotivité. Anthropomorphiser actuellement les SIA ne serait pas une raison pertinente pour parler de la responsabilité autonome à leur égard. Une responsabilité sans sujets de responsabilité n'est pas envisageable. Les artefacts, bien qu'ils soient des IA fortes, ne peuvent jouir d'une autonomie morale au point que l'homme se résout à leur confier la liberté de décision et d'en être des marionnettes à leur tour. Pour le philosophe de la responsabilité « prétendre que, outre l'exécution des tâches, l'automate puisse lui-même devenir vivant, acquérir une âme et qu'il puisse désormais, en vertu de sa propre volonté, nous donner du fil à retordre, n'est que pure spéculation (…)» (H. Jonas, 2017, pp. 101-102). Une telle position de Jonas se laisse cerner aisément à partir des fondements et des conditions de possibilité de la responsabilité esquissés précédemment.

L'IA ne jouit rigoureusement du statut de sujet juridique ni celui du sujet moral pour être un agent responsable. Un véhicule, un robot-tueur ou "robot-médecin", prétendument autonome souffre d'une absence de vie subjective comprise comme vie consciente, téléologique, libre et volontariste. Dans une perspective jonassienne, la responsabilité est siégée dans un sujet conscient et libre, puisqu' « un devoir est contenu très concrètement dans l'être de l'homme existant ; sa qualité de sujet capable de causalité comme telle entraîne l'obligation objective sous forme de la responsabilité externe » (Jonas, 1990, p. 194). Chez Jonas le pouvoir dont dispose l'homme lui impose une responsabilité. En faisant le parallèle avec

une telle considération, on pourrait être amené à dire que le pouvoir des IA est une raison pour parler de la responsabilité des IA. Mais le pouvoir, s'il est une condition nécessaire, ne serait pas une condition suffisante, car « Le pouvoir devient objectivement responsable pour ce qui lui est confié de cette manière et il est engagé affectivement par la prise de parti du sentiment de responsabilité : dans le sentiment ce qui oblige découvre son lien avec la volonté subjective » (Jonas, 1990, p. 182). L'affectivité et la volonté subjective sont absentes dans les SIA.

La question de la responsabilité en lien avec les IA se situe en dehors d'elles. Il s'agit d'inviter les concepteurs, les entreprises et les développeurs à plus de responsabilité. Le souhait d'Asimov de voir des robots capables de sens éthique reste une entreprise irréalisable tant de la complexité des questions éthiques. Le succès et les promesses des IA, comme tout autre invention technique, entrainent dans la perspective jonassienne une responsabilité humaine. Il résulte qu'il est de notre responsabilité de conduire l'avenir de l'IA selon nos échelles de valeurs. Il est question aujourd'hui de penser des IA qui doivent assister l'homme et non le remplacer sur des questions hautement éthiques.

Conclusion

Sur le sujet de la responsabilité de l'IA, il convient de retenir avec Jonas que l'IA ne peut jouir du statut de sujet moral ni de celui d'un objet de responsabilité. Dès lors, l'humanité ne peut déléguer son sens de responsabilité morale ou même juridique aux IA. Pour éviter le *nihilisme* éthique qui accompagne les progrès techniques de façon générale et en particulier les IA autonomes, Jonas estime que la responsabilité de l'homme devient encore plus grande qu'elle ne l'était auparavant. Si doter une IA d'une conscience morale serait très compliqué, cela revient à dire que seul l'homme doit répondre de ses actes. Une responsabilité qui se situe en amont et en aval de ses inventions, en l'occurrence le développement et l'utilisation des IA. Les IA accroissent le sens moral de l'homme, une responsabilité qui se charge de faire perpétuer la possibilité de responsabilité humaine. Des artefacts dénués de volonté, de liberté, de transcendance, en un mot d'*agencité*, ne peuvent assumer une responsabilité morale sur des

sujets humains. Seuls des sujets humains, au nom de l'enracinement ontique de la responsabilité, peuvent engager des débats sur la responsabilité historique qu'ils ont à assumer devant leurs artefacts. Les IA ne peuvent peut-être que les assister et non assumer toutes seules des enjeux moraux, sociaux, politiques, d'un agir. L'appauvrissement du sens de la responsabilité humaine nous conduira dans un monde dominé par le nihilisme radical, ici l'impossibilité de sens de l'existence et du monde, l'incapacité morale, et l'impossibilité de justifier ou de légitimer quoi que ce soit. Les conditions de possibilité de la responsabilité chez Jonas pourraient nous aider à éviter un tel nihil qui nous guette avec les IA.

Bibliographie

Bostrom Nick, (2017), *Super intelligence*, Malakoff, Dunod.

Desbiolles Jean-Philippe et Colombet Grégoire, (2023), *Hunain ou IA ? Qui décidera le futur ? Défis et opportunités d'un monde où l'IA nous dépasse*, Malakoff, Dunod.

Dreyfus L. Hubert, (1984), *Intelligence artificielle : Mythes et limites*, trad. de Rose-Marie Vassallo-Villaneau avec le concours de Daniel Andler, Paris, Flammarion.

Goffi J. Yves, (1993), « Communauté éthique et communauté politique », in Gilbert Hottois (éd.), *Aux fondements d'une éthique contemporaine. H. Jonas et H. T. Engelhardt*, Paris, Vrin pp. 205-222.

Jonas Hans, (1990), *Le Principe Responsabilité. Une éthique pour la civilisation technologique*, trad. de Jean Greisch, Paris, Cerf.

Jonas Hans, (1994), *Le Concept de Dieu après Auschwitz. Une voie juive. Suivi d'un essai de Catherine Chalier,* trad. de Philippe Ivernel, Paris, Payot & Rivages.

Jonas Hans, (2000), *Le Phénomène de la vie. Vers une biologie philosophique*, trad. de trad. de Danielle Lories, Bruxelles, Deboeck Université Bruxelles.

Jonas Hans, (2000), *Puissance et impuissance de la subjectivité*, trad. de Christian Arnsperger, Revue et présentée par Nathalie Frogneux, Paris, Cerf.

Jonas Hans, (2017), *Une éthique pour la nature*, trad. de Sylvie Courtine-Debamy, Paris, Flammarion.

Neerdael Dorian, (2018), « Singuralirité », *L'humain et ses préfixes. Encyclopédie du transhumanisme et posthumanisme*, Gilbert Hottois, Jean-Noël Missa et Laurent Perbal (dir.), Paris, J. Vrin, pp.438-442.

Pommier Éric, (2013), *Jonas*, Paris, Les Belles Lettres.

Russell Stuart et Norvig Peter, (2021) *Intelligence artificielle. Une approche moderne*, 4e édition, trad. de Claire Cadet, Laurent Miclet et Fabrice Popineau, Revision Claire Cadet, Person.

Vannieuwenhuyze Aurélien (2019), *Intelligence artificielle vulgarisée. La Machine Learning et le Deep Learning par la pratique*, ENI.

Weyembergh Maurice, (1993), « Before and after virtue », *Aux fondements d'une éthique contemporaine. H. Jonas et H. T. Engelhardt*, Gilbert Hottois (éd.), Paris, éd. Vrin pp.181-204.