Giornale di filosofia, vol. 2/2021 • ISSN 1827-5834 • DOI 10.7413/1827-5834011

Alexandre Bretel[*]

Quelle responsabilité adopter avec l'éthique de l'IA? L'apport des penseurs de la technique allemands du XX[e] siècle

Dans les rapports et les événements en éthique de l'intelligence artificielle (IA), la notion de responsabilité tend à remplacer celle d'éthique, les deux termes étant utilisés de manière synonyme. Au-delà de l'effet de communication, il est indispensable d'interroger le sens même qui est donné à cette notion. La responsabilité est définie au sens large comme l'obligation morale, voire légale, de se porter garant de ses actions ou de celles des autres. La notion de responsabilité peut aussi bien être pensée dans sa forme élémentaire, son institutionnalisation juridique, sa dimension éthique ou encore politique. L'angle d'approche qui sera abordé au cours de cette étude sera celui de la philosophie de la technique, et plus particulièrement celui des penseurs allemands du XXe siècle qui ont conceptualisé la notion de responsabilité.

Des penseurs tels que Hans Jonas, Günther Anders ou Hannah Arendt, ont introduit dans la philosophie de la technique des concepts qui ont révolutionné la discipline. Bien que ces penseurs n'aient pas traité directement de l'éthique de l'IA en tant que telle, à quelques exceptions près, comme par exemple la réflexion d'Hans Jonas sur les conséquences matérielles et morales de l'automatisation, ce sont leurs réflexions sur l'objet technique de manière générique qui sont appliquées dans cette étude. Leurs travaux sont d'une étonnante pertinence pour améliorer la compréhension de notre rapport à cet objet technique, notamment du point de vue de notre responsabilité. De plus, la plupart de ces penseurs prétendaient fonder des principes éthiques dépassant le cadre de leur époque et de leur objet d'étude, et devenant d'autant plus nécessaires avec le progrès constant de la technique.

Hans Jonas a ainsi développé le principe de responsabilité, stipulant la nécessité de la préservation d'une l'humanité authentique sur Terre, en utilisant l'heuristique de la peur comme méthode pour identifier les menaces au bien-être et à la survie de l'humanité. Non pas une peur

[*] Doctorant en philosophie. Université Grenoble Alpes

guidée uniquement par nos instincts ou nos premières impressions, susceptible d'être malléable au sensationnalisme et pouvant se révéler contre-productive, mais une démarche réfléchie et exemplifiée. Nous devrions donc écouter à la fois notre instinct mais aussi notre raison pour déterminer la peur la plus appropriée. Paradoxalement, nous devons trouver le courage d'avoir peur. Il s'est particulièrement intéressé aux risques liés aux modifications génétiques et a été un précurseur de la pensée écologique. C'est le principe responsabilité qui est à l'origine du principe de précaution, et qui s'oppose au principe espérance formulé par Ernst Bloch.

Günther Anders se demandait comment vivre à notre époque, qu'il décrit comme une période pré-apocalyptique, dans laquelle notre existence reste suspendue à la menace créée par la technologie. Il a écrit dans un contexte de guerre froide où le risque d'holocauste nucléaire était omniprésent. Il estime que la peur n'est pas seulement une attitude légitime, mais qu'elle est indispensable pour adopter une position morale. Nous vivrions dans un "décalage prométhéen", où nos capacités de production technique dépassent nos capacités de représentation, avec le risque d'une menace que nous ne pouvons même plus imaginer nous-mêmes. Quant à Hannah Arendt, elle a écrit sur la banalité du mal, dans le contexte de la Seconde Guerre mondiale, et se demande comment penser la responsabilité individuelle et collective pour éviter les pires dérives idéologiques, renforcées par la puissance de la technologie.

Au cours de notre étude, des comparaisons ont été effectuées entre les penseurs susmentionnés. La question de l'autonomie de l'agir technique humain pour Hans Jonas est ainsi à mettre parallèle de la conception autonomiste de la technique pour Günther Anders. Bien que le début de l'ouvrage *Principe responsabilité* s'ouvre avec la figure classique de Prométhée de coloration andersienne, celle-ci diffère de sa conception antécédente en se positionnant d'un point de vue éthique, en admettant que des règles volontairement contraignantes permettent de conditionner le comportement de la recherche et développement à des principes et des pratiques établies au préalable, malgré le fait que les humains soient devenus "plus grand qu'eux-mêmes".

Appliquée à l'éthique de l'IA, cette réflexion permettrait de donner du crédit, du moins sur le principe, aux initiatives visant à orienter certaines pratiques via des chartes éthiques, mais avec la prise de conscience d'un "décalage prométhéen" entre les capacités de calcul et de rapidité des ingénieurs comparées à celle de l'intelligence artificielle, et dont il convient de prendre la mesure. Hannah Arendt reprend également la thèse du décalage prométhéen, en supposant que nous ne "soyons plus jamais capables de comprendre, c'est-à-dire de penser et d'exprimer, les choses que nous sommes cependant capables de faire". Les réflexions des

élèves d'Heidegger sur la philosophie de la technique sont loin d'avoir pu être complètement explicitées et adaptées aux enjeux des technologies contemporaines, notamment à l'éthique de l'intelligence artificielle. Même s'il convient de rappeler qu'une partie significative de la pensée écologique et environnementale repose à l'origine, directement ou non, sur le dialogue entre ces auteurs et leur conception sur la manière de préserver ce qui doit l'être.

Leurs interactions, leurs complicités et leurs différents rendent ces cheminements d'autant plus passionnants. L'augmentation des capacités d'action conférée par la technoscience oblige à la formulation d'une nouvelle forme de responsabilité. Classiquement, la responsabilité est entendue comme l'obligation d'assumer son acte, par exemple en expiant, dans le cas d'une faute, ou en réparant les dommages le cas échéant. Dans le cadre de la civilisation technologique, la responsabilité est à comprendre comme la sollicitude que doit avoir un individu pour une chose ou une personne vulnérable si elle lui est confiée.

Dans *Principe responsabilité*[1], Jonas renvoie à *L'Obsolescence de l'homme*, qui traite des "possibles conséquences destructrices"[2] de la technique moderne. Le *Principe responsabilité* se réapproprie la thèse andersienne de "décalage prométhéen" entre notre faculté de produire et notre imagination. Il explique "la grandeur excessive" de notre "pouvoir" par "un excès de notre pouvoir de faire sur notre pouvoir de prévoir et notre pouvoir d'évaluer et de juger"[3]. Mais alors qu'Anders observe en psychologue, voire en psychiatre, ses Prométhées honteux qu'il décrit comme des êtres souffrants d'un complexe d'infériorité vis-à-vis de leurs productions, c'est en éthicien que Jonas considère, lui, son Prométhée capable de suivre, des règles éthiques "par des entraves librement consenties"[4]. Hannah Arendt a également repris la thèse andersienne du "décalage" prométhéen: "il se pourrait, créatures terrestres qui avons commencé d'agir en habitants de l'univers, que nous ne soyons plus jamais capables de comprendre, c'est-à-dire de penser et d'exprimer, les choses que nous sommes cependant capables de faire"[5].

Anders pense, comme Jacques Ellul, en l'autonomie de la technique. C'est-à-dire que la technique moderne prive peu à peu l'homme de sa liberté. Elle commence par le déclarer obsolète, tout en projetant sur le long terme de le liquider. Elle traite l'homme non seulement comme un

[1] H. Jonas, *Le Principe responsabilité*, Flammarion, Paris 1979, p. 149.
[2] G. Anders, *L'obsolescence de l'homme: Tome 1, Sur l'âme à l'époque de la deuxième révolution industrielle*, Ivrea, Paris 1956, p. 419.
[3] H. Jonas, *Le Principe responsabilité*, cit., p. 58.
[4] Ivi, p. 15.
[5] H. Arendt, *Condition de l'homme moderne*, Pocket, Paris 1958, p. 36.

simple moyen, mais carrément comme une matière première. La seule liberté qui reste à l'homme, chez Anders, est celle de résister à la technique qui travaille à l'évincer progressivement de la fonction de sujet de l'histoire pour l'y remplacer. En conséquence, la temporalité dans laquelle se trouve l'humanité s'en trouve bouleversée. Le futur n'est plus ouvert, mais prend la forme d'un délai, c'est-à-dire d'un sursis. L'action n'y est plus libre, au sens kantien, mais prend la forme d'une résistance, dont l'objectif est de retarder l'échéance autant que possible. Appliquée à l'éthique de l'intelligence artificielle, cette thèse provocante, jugée à l'aune des discours sur la responsabilité et la maîtrise, donne un point de vue dissonant, qui rappelle à tout le moins, que nous ne sommes pas complètement maîtres des processus techniques que nous enclenchons, voire pas du tout, du moins selon Anders.

À l'inverse, pour Jonas, l'histoire de la technique n'a pas eu d'effet irréversible sur l'homme. L'homme des sociétés industrielles serait resté identique au libre sujet kantien. Il reste donc le sujet à part entière de ses "actions techniques"[6]. Cet "homme", ce n'est "non pas vous ou moi: c'est l'agent collectif, non l'agent individuel"[7]. L'autonomie a ici très précisément la valeur ou le statut théorique d'une hypothèse ou d'un principe. Elle permet de parler abstraitement d'une libre action technique sans en rendre aucun agent véritablement responsable. Si Jonas reprend l'analyse andersienne du "décalage prométhéen", ce n'est donc pas pour la développer dans le sens d'une thèse sur l'autonomie de la technique, mais uniquement afin de confirmer et renforcer la nécessité d'une éthique.

Il y a donc un sens nouveau, donné à la même notion, du décalage prométhéen, repensé par Jonas, à savoir la possibilité de formuler une éthique de l'intelligence artificielle qui tienne compte pour le programmeur de la supériorité de calcul de la machine sur ses propres capacités, et donc sur son rapport, parfois inconscient, de positionnement par rapport à celle-ci, pour en formuler une relation éthique. Notons au passage que c'est la notion jonassienne de principe responsabilité qui est la plus ouverte à l'expérimentation, comparée à la conception andersienne de principe de précaution. L'opposition entre principe responsabilité et principe de précaution, ayant déjà été observée dans le positionnement respectif de Jonas et d'Anders sur la question nucléaire, mérite d'être soulignée en ce qu'elle constitue une opposition peut-être plus courante que l'opposition avec le principe espérance.

Pour Jonas, toutes les éthiques du passé auront été des "éthiques de la simultanéité"[8]. Elles avaient pour objet l'action elle-même au moment

[6] H. Jonas, *Le Principe responsabilité*, cit., p. 65.
[7] Ivi, p. 37.
[8] Ivi, p. 43.

même où elle a lieu et non ses effets à long terme[9]. Seules trois éthiques auront fait exception à cette règle: "l'éthique de l'accomplissement dans l'au-delà"[10], celle de "la responsabilité de l'homme politique dans l'avenir"[11] et celle de "l'utopie moderne"[12]. Aujourd'hui, c'est une éthique du futur qu'il faudrait privilégier. Pour Jonas, comme pour Anders, la question du futur, c'est d'abord et surtout la nécessité de l'existence des générations futures. Selon Jonas, "nous n'avons pas le droit de choisir le non-être des générations futures à cause de l'être de la génération actuelle, ni même le droit de prendre le risque de leur non-être"[13]. Nous avons d'autant moins ce droit, soutient Anders, que l'humanité a maintenant réellement la possibilité de se détruire elle-même[14]. La nécessité de l'existence des générations futures semble finalement jouer, chez Jonas, comme chez Anders, le rôle d'un postulat de la raison pratique qui doit être posé avant tout développement éthique.

Appliqué à l'éthique de l'intelligence artificielle, cette réflexion pousse à se projeter davantage dans les objectifs à long terme poursuivis par les projets numériques. Par exemple, en évitant de développer un algorithme dédié à des applications dont les conséquences pourraient être néfastes pour l'humanité. Ainsi, les programmeurs d'algorithme militaires doivent toujours penser à proportionner le développement de leur projet à un objectif précis et délimité. Il s'agit de garder à l'esprit cet impératif, comme un soubassement à toute action responsable. Le problème est dès lors de savoir comment ce postulat détermine l'action responsable chez Jonas et chez Anders. Force est de constater qu'ils n'ont pas l'un et l'autre la même conception du futur. Il est ouvert pour Jonas et fermé pour Anders. C'est pour cette raison que le postulat de la nécessité de l'existence des générations futures n'a pas le même sens chez l'un et chez l'autre. Leurs conceptions respectives du futur déterminent donc les horizons dans lesquels s'inscrivent leurs concepts respectifs de l'action responsable.

Ces conceptions conduiraient les programmeurs à assumer leur autonomie et à se considérer comme responsables, au moins éthiquement, des actions qu'ils entreprennent, afin que leur conscience soit engagée dans les conséquences de leurs actions. Au cœur du *Principe responsabilité*, on trouve la méthode de l'heuristique de la peur. Celle-ci postule que la peur "fait essentiellement partie de la responsabilité"[15], bien qu'il

[9] Ivi, p. 40.
[10] Ivi, p. 43.
[11] Ivi, p. 45.
[12] Ivi, p. 47.
[13] Ivi, p. 40.
[14] G. Anders, *L'obsolescence de l'homme,* cit., p. 269.
[15] H. Jonas, *Le Principe responsabilité*, cit., p. 421.

faille entendre la peur au sens philosophique plus qu'au sens émotionnel. Néanmoins, cette méthode ne permet pas de déterminer à l'avance si une catastrophe aura lieu ou non dans le futur. Elle permet seulement d'affirmer qu'une catastrophe pourrait avoir lieu si l'on agissait de telle ou telle façon. Il ne s'agit pas de faire des prévisions, mais d'examiner des hypothèses. L'heuristique de la peur revient ainsi à évaluer les conséquences des actions que nous pouvons accomplir dans le présent du point de vue du futur.

Il s'agit donc, du point de vue de l'éthique artificielle, d'adopter une posture d'heuristique de la peur, c'est-à-dire d'avoir peur, au sens philosophique, des conséquences potentielles de son action, tout en se projetant du point de vue du futur des conséquences éthiques. Une autre différence importante entre les pensées d'Anders et de Jonas, avec le *Principe responsabilité* consiste en ceci que Jonas entreprend de fonder une éthique alors qu'Anders, pour sa part, s'y refuse. Jonas dit vouloir fonder une éthique pour la civilisation technologique, alors qu'Anders, pour qui "les éthiques religieuses et philosophiques qui ont été jusqu'ici proposées dans l'histoire sont devenues obsolètes", et qui en outre, persuadé que vouloir fonder une nouvelle éthique est une "entreprise utopique", ne se considère pas comme un éthicien mais comme un moraliste. Par contraste, la singularité de la philosophie de Jonas tient au fait qu'elle renoue avec un geste classique, voire antique, celui qui consiste à fonder l'éthique sur la métaphysique.

Ces positions sont en contraste total avec Anders, pour qui la fondation d'une telle métaphysique est rendue impossible, à la fois par refus de la religion, autrement dit de la métaphysique qui la soutient, ainsi que par refus de l'anthropocentrisme et du biocentrisme. C'est la technique qui, malgré nous, énonce ses commandements. En tant que moraliste, Anders s'adapte aux mœurs de son époque, contrairement à l'éthique, qui formule des impératifs universels. Il y a donc deux approches possibles pour les personnes impliquées dans les projets d'IA, soit partir d'un point de vue éthique, c'est-à-dire selon des principes universels et clairement établis, soit adopter une posture morale qui tienne compte de l'évolution des mœurs. Une réflexion sur la morale artificielle, à distinguer de l'éthique artificielle, serait intéressante pour s'adapter aux mœurs d'une époque donnée.

La question de la responsabilité est une question transversale de première importance dans l'œuvre d'Hannah Arendt. Ainsi, par exemple, l'une des questions les plus graves soulevées par sa pensée est celle de responsabilité des agents des criminels nazis allemands, considérés à la fois comme banals et monstrueux. Il faut à la fois éviter de diluer leur responsabilité tout en reconnaissant qu'ils sont intégrés comme les rouages d'un système dans une machine de mort bureaucratique. C'est

le développement prodigieux de nos capacités et de nos moyens d'action sur le monde environnant au cours des cinq derniers siècles qui a modifié à la fois le sens et la valeur de nos actions, et donc corrélativement de notre responsabilité.

Comment tenir l'individu pour responsable de ses actes dès lors que ceux-ci s'intègrent dans une longue chaîne de processus techniques, en amont et en aval de son action, lesquels aboutissent à des conséquences dont il n'a pas connaissance et sur lesquelles il n'a pas de prise? Pour le dire autrement, et le point de vue d'Arendt rejoint en cela celui de Jonas et d'Anders, la modernité scientifique et technique a suscité un divorce entre la pensée et l'action. Et encore plus, le schème dominant de la connaissance moderne, à savoir l'expérimentation scientifique, a consacré une confusion totale entre la pensée et l'action au profit de cette dernière. La pensée et ses résultats sont soumis au verdict produit par le protocole expérimental.

La technoscience contemporaine est en mesure de déclencher des processus de transformation du monde que l'homme a certes engendrés, mais sur lesquels celui-ci risque ensuite de ne plus avoir aucune prise. Pire encore, les processus technoscientifiques déclenchés par l'homme menacent de se retourner contre lui, comme l'avait annoncé Anders avec la menace nucléaire. Il y a donc un processus de déresponsabilisation à l'œuvre. L'éthique de l'intelligence artificielle doit donc tenir compte du fait que la responsabilité des programmeurs peut être diluée, par exemple la responsabilité des programmeurs du régime chinois qui participent à la surveillance de masse au moyen d'algorithmes. Il faut donc réfléchir à la fondation d'une responsabilité qui concilie à la fois la responsabilité individuelle et collective, dont on peut concevoir qu'elle serait une "responsabilité interpersonnelle".

Il faut donc repenser l'action, et donc la responsabilité des hommes d'aujourd'hui. La solution passe par la responsabilité politique, à distinguer de la responsabilité morale et philosophique. Le problème d'Arendt est ainsi de tenter de découvrir un nouveau point de vue sur l'action et la responsabilité humaine qui tienne compte des transformations du monde moderne. De fait, on peut déduire de sa conception de l'agir comme agir au sein du réseau des relations humaines, une nouvelle conception de la responsabilité susceptible de s'étendre au-delà de la responsabilité strictement individuelle, tant morale que juridique, pour prendre en considération la dimension relationnelle de la notion de responsabilité.

En cela, les propositions de Jonas pour remédier à cette crise sont foncièrement différentes de celles d'Arendt. En effet, l'éthique jonassienne de la responsabilité est une éthique de la préservation de la vie, au sens biologique du terme, ce qui la distingue radicalement de l'éthique arendtienne de la responsabilité, laquelle est une éthique de la préservation du

monde en tant que domaine de l'interhumain. Dans le cadre de l'éthique artificielle, il faut donc penser une éthique qui tienne compte de la responsabilité à la fois individuelle et collective, mais également interpersonnelle, qui reste à définir, pour s'adapter à la fois aux évolutions techniques mais aussi pour faire progresser la notion même de responsabilité.

Bibliographie

Anders G., *L'obsolescence de l'homme: Tome 1, Sur l'âme à l'époque de la deuxième révolution industrielle*, Ivrea, Paris 1956.
Id., *L'obsolescence de l'homme: Tome 2, Sur la destruction de la vie à l'époque de la troisième révolution industrielle*, Fario, Paris 1980.
Arendt H., *Les origines du totalitarisme*, Gallimard, Paris 1951.
Id., *Condition de l'homme moderne*, Pocket, Paris 1958.
Id., *Responsabilité et jugement*, Payot, Paris 1950-1973.
Béranger J., *Comment évaluer l'éthique d'un algorithme?*, in "The Conversation", 2018.
Bodet M., *La notion de "responsabilité" chez les penseurs allemands et français de la durabilité*, in "ERIED (Équipe de Recherche et d'Etudes Interdisciplinaires sur la Durabilité)", 2013.
Boujemaa N., *De l'éthique de l'intelligence artificielle*, in "AI Paris Interview", 2019.
David, C. & Röpcke D*., Günther Anders, Hans Jonas et les antinomies de l'écologie politique,* in "Ecologie & Politique", 2004, pp. 193-213.
Gilbert M., *Faire la morale aux robots: une introduction à l'éthique de l'intelligence artificielle*, Flammarion, Paris 2021.
Jonas H., *Le Principe responsabilité*, Flammarion, Paris 1979.
Poizat J.-C., *Assumer l'humanité, Hannah Arendt: la responsabilité face à la pluralité de Gérôme Truc*, in "Le Philosophoire", 2009, pp. 177-188.
Vaissière T., *L'éthique de responsabilité chez Hans Jonas à l'épreuve du droit international de l'environnement*, in "Revue interdisciplinaire d'études juridiques", 1999, pp. 135-199.