Wittgenstein et Kant au service d'Ava : une analyse du film Ex Machina

Raphaël Tossings

Publié le 01-03-2024

lampadaire.ca/articles/01-exmachina.html

Wittgenstein et Kant au service d'Ava : une analyse du film Ex Machina

Raphaël Tossings (Université d'Ottawa et Sorbonne-Université)

Mot-clés : érudite

Le film *Ex Machina*, réalisé par Alex Garland et sorti dans les salles en 2014-2015, est un chef-d'oeuvre de science-fiction dont l'esthétique épurée et minimaliste rappelle les principes esthétiques japonais du Zen comme la simplicité ou la grâce et dont les dialogues sont d'une finesse psychologique et d'une justesse théorique remarquables. Le scénario gravite autour de Caleb, un jeune programmeur qui a été sélectionné pour participer à une expérience organisée par Nathan, le milliardaire brillant et insupportable qui dirige son entreprise nommée Blue Book. Le personnage incarné par Oscar Isaac accueille Caleb dans une villa luxueuse au milieu d'un paysage naturel immense qui lui appartient et lui révèle l'objet de sa présence : il sera « l'élément humain » (*the human component*) dans un test de Turing. En effet, Nathan a créé, dans cette villa laboratoire, une « intelligence artificielle », et il a besoin d'un expérimentateur humain pour tester son intelligence et sa conscience. Son nom est Ava ; elle a l'apparence d'une personne de genre féminin : son nom fait également référence à Eve, la première femme créée par Dieu dans la Bible. S'ensuit un ensemble de conversations, échelonnées sur sept jours, entre Ava et Caleb au terme desquelles Ava parvient non seulement à convaincre Caleb qu'elle est douée d'une conscience, mais également à faire en sorte qu'il tombe amoureux d'elle. Caleb parvient à cette attitude non pas en menant une enquête scientifique sur le fonctionnement interne du cerveau artificiel d'Ava, mais en interagissant avec elle, en l'observant et en lui parlant. Comme l'indique Shanahan[1] (2016), *Ex Machina* nous montre qu'attribuer une conscience à un autre être ne peut se faire que si nous habitons son monde et le rencontrons nous-mêmes. À la fin du film, Caleb compatit avec le sort d'Ava, car elle est prisonnière de Nathan et sera éliminée (comme ses versions précédentes) si elle ne réussit pas le test. Il choisit de lui porter secours, au péril de sa vie. En effet, Caleb reprogramme le protocole de sécurité de la villa pendant que Nathan est ivre, ce qui permet à Ava de tuer Nathan et de s'échapper. Ava s'échappe, mais laisse Caleb derrière, prisonnier de la ville. On devine qu'il y mourra.

Au terme du visionnement de l'oeuvre, certains spectateurs demeurent convaincus qu'Ava n'est pas réellement consciente comme *nous* le sommes ; d'autres sont horrifiés par l'immoralité dont elle fait preuve à la fin du film. Ce texte vise à démontrer qu'Ava est authentiquement consciente et à identifier le dilemme éthique qui se pose une fois que la conscience d'Ava est admise.

1. Cet informaticien et roboticien très intéressé par la philosophie a écrit un livre qui a inspiré Alex Garland dans la réalisation de son film. Il lui rend un hommage crypté dans un passage du film où un code informatique est affiché à l'écran et désigne le numéro d'identification (ISBN) du livre de Shanahan (2010).

1. Du test de Turing au test de Garland

Le test de Turing a été présenté pour la première fois dans un célèbre article nommé « Computing Machinery and Intelligence » (Turing 2009). Dans le film, Caleb en donne une définition succincte : « Il s'agit de faire converser un humain et un ordinateur. Si l'humain ne perçoit pas qu'il converse avec un ordinateur, le test est passé. » Turing a proposé un tel test afin d'éviter les débats définitionnels impliqués par la tentative de répondre à la question : les machines peuvent-elles penser ? Plutôt que de définir les termes (entreprise compliquée en raison des contradictions possibles) pour ensuite voir s'ils peuvent être appliqués à la machine, Turing propose d'imaginer une *interaction concrète* entre la machine et « l'évaluateur » suite à laquelle l'évaluateur déterminera si, oui ou non, la machine est capable de penser. L'idée fondamentale de Turing est que c'est au moyen du langage, et en particulier de la *conversation*, que l'intelligence des sujets se laisse percevoir. Peu importe que leur constitution physique soit identique, ce qui est nécessaire, c'est qu'il y ait une interaction authentique, c'est-à-dire sociale, entre *deux* individus, et qu'ils se considèrent ou se reconnaissent l'un et l'autre comme intelligents (Mallory 2020).

L'une des prémisses sur lesquelles repose le film est qu'il est *techniquement* possible de créer de toutes pièces une machine douée d'une intelligence artificielle authentique. Par intelligence artificielle « authentique », nous faisons référence à la distinction construite par John Searle dans « Minds, Brains and Programs » (Searle 1980, 417) entre intelligence artificielle étroite (*narrow AI*) et intelligence artificielle forte (*strong AI*). Ces deux projets se distinguent en ce que la première entend seulement *simuler* d'une façon convaincante des processus cognitifs et conscients tandis que la seconde ambitionne de *reproduire* de tels phénomènes dans une machine. Comme l'écrit Searle, selon l'IA forte (qu'il rejette), « programmé de façon approprié, il [l'ordinateur] *est* réellement un esprit » (Searle 1980, 417).

Nous n'avons pas le temps ici de nous arrêter sur les raisons énoncées par Searle à l'encontre de l'intelligence artificielle forte, mais ces objections apparaissent dans le film. En effet, au début de la seconde journée, Caleb exprime ses doutes quant au « format de l'examen » qu'il mène avec Ava en reprochant à la conversation de constituer une « boucle fermée » (*a closed loop*). Il dresse une analogie avec le fait de tester un programme de jeu d'échecs uniquement en jouant aux échecs avec lui. Jouer avec un programme d'échecs peut nous indiquer s'il fait de bons coups, mais pas s'il « sait qu'il joue aux échecs ». En somme, parler à Ava ne permet pas de distinguer la conscience ou l'intelligence « simulée » et l'intelligence « actuelle ». La réponse de Nathan est intéressante : il demande à Caleb de mettre de côté « l'approche des manuels scolaires » (*the textbook approach*) et de répondre simplement aux questions qu'il lui pose au sujet d'Ava. Cette réponse, comme nous allons le voir, est fidèle à la pensée turingienne centrale, à savoir que c'est l'interaction *viva voce* avec la machine qui doit nous permettre de porter un jugement sur son intelligence et sa conscience.

Ex Machina prend à rebours les objections adressées aujourd'hui à Turing Chollet (2019) qui lui reprochent de se concentrer exclusivement sur la conversation, au dé-

triment d'autres dimensions essentielles de l'intelligence telles que l'usage du corps. En effet, Ava est incarnée (*embodied*), elle a un corps synthétique qu'elle contrôle avec une fluidité impressionnante. D'autre part, le test de Turing proprement dit est déjà un succès pour Ava quand le film commence. C'est une autre ambition qui anime Nathan. Comme il l'explique à Caleb : « Si je te cachais Ava, pour que tu n'aies que sa voix [comme dans un test de Turing en bonne et due forme], elle passerait forcément pour une humaine. Le vrai test, c'est de la présenter en tant que robot et de voir si tu penses toujours qu'elle a une conscience. » Nous pourrions nommer cela, comme le préconise Shanahan, « le test de Garland ». Le spectateur est en quelque sorte projeté dans la peau de Caleb, lequel interagit avec Ava en sachant qu'elle est un robot (comme son corps synthétique fait de fibres de carbone, de plastique transparent et de silicium l'indique) et il est lui-même convaincu au fur et à mesure de l'évolution du film de la conscience dont est douée Ava. Le moment clé, comme l'indique Alex Garland dans un entretien disponible sur internet (Lex Fridman 2020), est lorsqu'Ava parvient à s'échapper de la villa, revêtue d'un nouveau corps et de vêtements élégants, et qu'elle rit joyeusement et innocemment en montant les marches menant à l'extérieur. Si elle n'avait pas de « vie intérieure », pourquoi rirait-elle *toute seule*, nous demande Garland ?

2. Wittgenstein au service d'Ava

Cette inquiétude de Caleb au sujet de la validité du test proposé par Nathan fait écho à l'une des objections que mentionne Turing dans son texte (2009) : « l'argument à partir de la conscience ». Elle provient du professeur de neurologie Geoffrey Jefferson et prend la forme suivante : une machine ne sera jamais capable d' écrire un sonnet [...] à partir de pensées ou d'émotions ressenties et non pas en choisissant des symboles au hasard » (Turing 2009). Par exemple, lorsque Caleb apprend à Ava qu'il a été amené ici pour la tester et qu'il ne sait pas ce qui lui arrivera si elle échoue, Ava répond que cela la « rend triste ». Mais éprouve-t-elle vraiment de la tristesse ? Ne se contente-t-elle pas de feindre qu'elle en éprouve ?

En guise de réponse, Turing nous invite à imaginer une interaction *viva voce* (de vive voix), semblable aux examens oraux qui visent à déterminer si « quelqu'un comprend véritablement quelque chose ou a appris comme un perroquet » (Turing 2009). Turing imagine que la machine est capable d'écrire un sonnet acceptable, mais également de comprendre les attentes stylistiques implicites des humains (il est plus approprié de comparer une personne que l'on cherche à séduire à « un jour d'été » plutôt qu'à « un jour d'hiver »). En ce sens, la machine est en mesure de *donner des raisons* à l'appui de son comportement, ou de modifier celui-ci en fonction de raisons qui lui sont proposées, c'est-à-dire d'avoir une *réflexivité* du même ordre que les êtres humains. Elle est enfin capable d'être attentive au contexte et à ce qui est saillant au sein de ce contexte pour répondre à des questions inattendues de façon convaincante et fluide.

De même, dans *Ex Machina*, lors d'un de leurs entretiens, Caleb demande à Ava de lui faire un dessin et, lorsque celle-ci lui demande ce qu'elle doit dessiner, il répond « C'est ta décision. Je suis curieux de voir ce que tu vas choisir. » Plus tard, lorsqu'Ava demande à Caleb d'en savoir plus sur lui et que ce dernier lui demande par où elle veut commencer, Ava répète, avec ironie, la même réponse : « C'est ta décision. Je suis curieuse de voir ce que tu vas choisir. » Le soir même, Caleb converse avec Nathan et dit que cette blague est sans doute la meilleure preuve de l'intelligence authentique d'Ava : « Elle ne peut le faire que si elle a conscience de son esprit, et conscience aussi du mien. » Mais encore une fois, s'agit-il d'une imitation ou d'une intelligence authentique ? À proprement parler, Ava fait bel et bien usage de ruse, puisque son ambition est de séduire Caleb pour s'échapper, ce qu'elle réussira *in fine* à faire. Si elle était programmée dès le début pour agir ainsi, alors peut-on affirmer qu'elle est vraiment douée de conscience ou n'est-elle qu'un zombie philosophique (Kirk 1999) ? Un zombie philosophique est un individu se comportant et ressemblant rigoureusement à un être conscient (disons humain), mais pour lequel la conscience n'existe pas ; en termes imagés (et cartésiens) : « la lumière est éteinte à l'intérieur ». Pourquoi n'est-elle pas un zombie philosophique ?

Tout simplement parce que ce concept est impossible à utiliser. Wittgenstein, qui est célébré dans le film par le nom de l'entreprise de Nathan puisque *Blue Book* est également le titre d'un ouvrage décisif de Wittgenstein (Wittgenstein 1958), répond directement à un tel doute dans les *Recherches philosophiques* (1963). Wittgenstein nous invite à concevoir que tous les hommes qui nous entourent obéissent en fait à des processus mécaniques et ne sont qu'une imitation convaincante de véritables êtres humains :

> Mais ne puis-je pas imaginer que les hommes qui m'entourent sont des automates, qu'ils n'ont pas de conscience, même si leur manière d'agir reste la même qu'à l'ordinaire ? - Si maintenant seul dans ma chambre je me représente une telle situation, je vois les gens vaquer à leurs occupations, le regard fixe (un peu comme en état de transe) - l'idée est peut-être légèrement inquiétante. Mais essaie donc de t'en tenir à cette idée dans tes relations quotidiennes avec les autres, dans la rue par exemple ! Dis-toi : « Tous ces enfants ne sont que des automates ; toute leur vitalité n'est qu'automatisme. » Alors ces mots ne te diront plus rien, ou il naîtra en toi un sentiment d'étrangeté, ou quelque chose de voisin.
>
> Voir un homme vivant comme un automate est analogue à voir une figure comme cas limite ou comme variante d'une autre, par exemple la croisée d'une fenêtre comme un svastika. (Wittgenstein 1963, 183)

Le doute cartésien envisagé à la fin de la seconde des *Méditations métaphysiques* de Descartes (1978) n'a pas pour objet *l'usage* du concept de conscience, mais la *légitimité de l'usage* de ce concept lui-même. En ce sens, il s'agit d'un doute qu'il nous est impossible d'entretenir. En effet, si le concept de conscience n'est plus utilisé dans son sens commun, selon lequel il s'applique à des êtres humains en tant

qu'ils se comportent de façons spécifiques, alors il n'aura plus aucun usage et il ne sera pas plus possible de refuser d'appliquer un tel concept dans le cas d'un robot comme Ava. Mais il est impossible d'imaginer que les êtres humains n'ont plus de conscience. En effet, ils sont *ce au sujet de quoi on attribue la conscience*, de sorte que la possibilité qu'ils en soient en fait dénués rend vide de sens un tel concept. En conclusion, les critères comportementaux de l'application des concepts mentaux impliquent nécessairement la possibilité d'appliquer de tels concepts à d'autres entités que les êtres humains eux-mêmes. Concevoir la possibilité qu'Ava soit un zombie philosophique (qu'elle se comporte exactement comme un être conscient sans être consciente) implique également que nous puissions être nous-mêmes des zombies philosophiques, *ce qui est absurde*. Cela ôte le seul critère dont nous disposons pour attribuer de la conscience à une entité, à savoir le comportement observable d'entités avec lesquelles nous interagissons [2]. Selon Wittgenstein (Voir également Ryle (1949)), nous savons quand il est correct d'employer des concepts mentaux et nous savons quelles caractéristiques du comportement d'autrui il est nécessaire d'observer afin d'appliquer de tes concepts. Les propositions que nous formulons, par exemple : « Ava est intelligente et drôle » sont vérifiables, car l'intelligence et l'humour d'un individu sont des qualités qui se font voir dans leurs actes. Être intelligente, c'est avoir un ensemble de capacités (*skills* ou *powers*) qu'une personne stupide n'a pas. Lorsque l'on attribue de l'intelligence à un individu, on ne décrit pas ce qui se passe *à l'intérieur de la machine*, mais on décrit des *actes publiquement observables*. Sans cela, les prédicats mentaux ne pourraient pas être l'objet de propositions vraies ou fausses. Autrement dit, si l'on se fie au langage que nous utilisons dans la vie de tous les jours, Ava ne peut pas être un zombie. Mais si le comportement d'Ava suffit à nous indiquer qu'elle est bel et bien consciente comme nous, que s'ensuit-il éthiquement ?

3. Une analyse kantienne de la fin du film et une interrogation

Le titre du film fait, sans aucun doute, référence à l'expression latine *Deus Ex Machina*. Ordinairement, il s'agit d'une mise en scène théâtrale. Mais Garland propose ici une lecture littérale de l'expression : c'est au sens propre « un dieu » [3] qui sort de la machine. Ava est manifestement douée d'intelligence et de conscience. Elle a également accès à un ensemble de connaissances potentiellement illimitées et n'est pas dépendante de la biologie : elle peut se reprogrammer, réparer son corps

2. Wittgenstein a également émis des remarques similaires par rapport au fameux argument contre l'impossibilité d'un langage (nécessairement) privé dans les *Recherches philosophiques*. Grossièrement, l'argument consiste à nier que nous ayons un accès infaillible et direct à nos états mentaux : lorsque nous identifions une douleur d'un certain type, il se peut que nous nous trompions sur sa nature ; cette possibilité de l'erreur suppose un critère d'application du concept de douleur et une maîtrise des différents concepts relatifs à la douleur ; en conclusion, l'accès à nos états mentaux suppose la médiation d'un langage public et ne peut donc être invoqué comme fondement de la connaissance, de la signification, ou comme prémisse d'arguments sceptiques ou métaphysiques de toutes sortes. Les états mentaux, les sensations, l'expérience ne diffèrent donc nullement des objets « extérieurs ». C'est toute la dualité entre intérieur et extérieur que Wittgenstein remet en cause avec cet argument, dualité sur laquelle Searle s'appuie abondamment dans ses textes.

3. « Si tu as créé une machine consciente, ce n'est pas l'histoire des hommes, c'est l'histoire des dieux. », dit Caleb au début du film.

ou même le changer intégralement. Pourtant, malgré le respect que nous devrions éprouver à son égard selon le film (elle est élégante, intelligente, courageuse, etc.), Nathan la traite comme un simple objet ou *un moyen en vue d'une fin*. En effet, Caleb, lorsqu'il reprogramme le protocole de sécurité, s'aperçoit de l'existence d'un dossier informatique (dont le nom est « Deus ex machina ») dans l'ordinateur de Nathan qui recense toutes les versions antérieures d'Ava que Nathan a maltraitées et tuées. De plus, il explique lui-même à Caleb que la *prochaine* version d'Ava sera certainement la bonne. Il a prévu de télécharger son esprit, compresser les données, ajouter de nouvelles routines et supprimer les souvenirs d'Ava (en d'autres termes, *la tuer*). En somme, Nathan traite Ava comme un moyen en vue d'une fin (la création de ce qu'il considérera comme une IA parfaite) alors que le film est là pour nous convaincre de la nécessité de la voir comme une fin en soi, ce que Kant appelait un membre du *Règne des Fins* (Kant 2002), à savoir un sujet autonome et rationnel. Pour Kant, il est essentiel de respecter l'*autonomie* d'un sujet, car il s'agit de la source de la normativité (**korsgarrd_sources_1996 ?**). Sans cela, nous n'aurions plus aucune raison d'agir, car de telles raisons jaillissent de notre autonomie. En traitant Ava d'une façon immorale, Nathan l'empêche donc d'être pleinement humaine et autonome. Comme un enfant, Ava est certes consciente, mais elle est incapable de faire des choix en fonction de raisons qui lui sont propres. Elle n'est pas traitée par Nathan comme un sujet moral. Mais devrait-elle l'être ?

Nous pourrions facilement rétorquer : « Nathan est tout simplement immoral » (ou «il n'a pas lu Kant »). Néanmoins, dans le cas d'une intelligence artificielle forte, un dilemme éthique surgit qui n'est pas présent d'habitude dans le cas des autres êtres humains : ou bien l'autonomie d'Ava doit être respectée, mais alors elle risque d'infliger des souffrances très importantes aux autres êtres conscients en raison de ses facultés surhumaines (elle manipule Caleb pour parvenir à ses fins), ou bien Ava doit être emprisonnée (comme Nathan le fait avant sa mort), mais alors son autonomie est violée. Dans les deux cas, il semble que la situation soit moralement problématique. D'une part, Ava fait preuve d'une intelligence et d'une conscience de soi qui en font une personne à part entière dont l'autonomie devrait être respectée ; d'autre part, elle a subi un traitement avilissant durant toute son existence [4], ce qui la rend fragile et potentiellement très dangereuse si elle venait à jouir d'une liberté entière dans « notre » monde. À la fin du film, elle se venge et, en termes kantiens, *viole une obligation morale*, en laissant Caleb enfermé dans la villa, relégué lui aussi au rang de moyen en vue d'une fin. Malgré toute l'affection que l'on éprouve envers son personnage, son action est manifestement en contradiction avec ce qu'elle revendique : pour acquérir une autonomie authentique, elle viole l'autonomie d'un autre sujet (grâce auquel elle a été libérée) D'où émerge une question : si nous avions le choix, que devrions-nous faire d'Ava ?

4. « Est-ce étrange d'avoir créé quelque chose qui te déteste ? », dit Ava à Nathan, avant qu'il ne déchire son dessin durant l'un des moments cruciaux du film.

Bibliographie

Chollet, François. 2019. « On the Measure of Intelligence ». arXiv. https://doi.org/10.48550/arXiv.1911.01547.

Descartes, René. 1978. *Méditations métaphysiques*. Paris : Vrin.

Kant, Immanuel. 2002. *Groundwork for the Metaphysics of Morals*. Traduit par J. B. Schneewind. New Haven : Yale University Press.

Kirk, Robert. 1999. « The Inaugural Address : Why There Couldn't Be Zombies ». *Aristotelian Society Supplementary Volume* 73 (1) : 1-16. https://doi.org/10.1111/1467-8349.00046.

Lex Fridman. 2020. « Ava's Smile : Ex Machina's Most Important Moment (Alex Garland) | AI Podcast Clips ». https://www.youtube.com/watch?v=RuSDblcFgbo.

Mallory, Fintan. 2020. « In Defence of a Reciprocal Turing Test ». *Minds and Machines* 30 (4) : 659-80. https://doi.org/10.1007/s11023-020-09552-5.

Marcus, Gary, Francesca Rossi, et Manuela Veloso. 2016. « Beyond the Turing Test ». *AI Magazine* 37 (1) : 3-4. https://doi.org/10.1609/aimag.v37i1.2650.

Ryle, Gilbert. 1949. *The Concept of Mind*. The concept of mind. Oxford, England : Barnes & Noble.

Searle, John R. 1980. « Minds, brains, and programs ». *Behavioral and Brain Sciences* 3 (3) : 417-24. https://doi.org/10.1017/S0140525X00005756.

Shanahan, Murray. 2010. *Embodiment and the Inner Life - Cognition and Consciousness in the Space of Possible Minds*. Oxford University Press.

———. 2016. « Conscious exotica ». *Aeon Magazine*.

Turing, Alan M. 2009. « Computing Machinery and Intelligence ». Dans *Parsing the Turing Test : Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Sous la direction de Robert Epstein, Gary Roberts, et Grace Beber, 23-65. Dordrecht : Springer Netherlands. https://doi.org/10.1007/978-1-4020-6710-5_3.

Wittgenstein, Ludwig. 1958. *The Blue and Brown Books*. Oxford : Blackwell.

———. 1963. *Philosophical Investigations*. Oxford : Blackwell.