Quand l'IA tue : 2001, l'odyssée de l'espace, ou le récit de la fin de l'espèce ?

Aurélien Portelli, Sébastien Travadel, Franck Guarnieri

Intelligence artificielle

Quand l'IA tue : *2001, l'odyssée de l'espace*, ou le récit de la fin de l'espèce ?

2018 marque le cinquantième anniversaire de la sortie au cinéma et en librairie de *2001, l'odyssée de l'espace*. Un cas d'affrontement mortel entre des astronautes et un super calculateur capable de reproduire la plupart des activités du cerveau humain. Ce scénario entre en résonance avec les craintes suscitées aujourd'hui par les développements récents et les promesses de l'intelligence artificielle. Au-delà d'une vision catastrophiste, *2001* offre cependant une perspective originale sur les risques liés à une superintelligence.

Poole et Bowman décident par précaution de déconnecter Hal 9000 - © Metro Goldwyn Mayer.

A l'été 1956 au Dartmouth College, à Hanover (New Hampshire), une vingtaine de scientifiques, dont Marvin Minsky, Herbert Simon et Allen Newell, conduisent le premier projet de recherche en intelligence artificielle (IA). Depuis, les machines intelligentes ont envahi notre quotidien et atteignent des performances impressionnantes. En 1996, Deep Blue d'IBM bat le champion du monde d'échecs Garry Kasparov. En 2011, Watson, toujours un produit d'IBM, gagne contre tous les champions du jeu télévisé *Jeopardy !* En 2015, DQN de DeepMind apprend à jouer à quarante-neuf jeux classiques sur la console Atari 2600 et atteint

des scores supérieurs aux champions humains. En 2016, AlphaGo de DeepMind bat au jeu de go Lee Sedol, l'un des meilleurs joueurs au monde. En décembre 2017, un algorithme de Google découvre l'exoplanète Kepler-90i parmi les masses de données recueillies par le satellite Kepler de la NASA.

Ces avancées technologiques émerveillent mais suscitent aussi les pires craintes. Le philosophe Nick Bostrom souligne notamment l'aveuglement des concepteurs : la plupart d'entre eux nie la possibilité d'une machine superintelligente, et ne veut pas admettre un risque existentiel « *qui menace d'entraîner l'extinction de la vie intelligente ayant pour origine la Terre* »[1]. L'autonomie croissante des machines, aux capacités surhumaines, aboutirait inéluctablement à une prise de contrôle violente et à l'anéantissement de l'humanité, jugée encombrante et inutile.

Or, *2001, l'odyssée de l'espace,* film-culte mettant en scène une IA meurtrière, amène à reformuler le danger de ces technologies. En effet, l'élimination de l'équipage ne procède pas dans 2001 de l'autonomisation du superordinateur Hal 9000 (dénommé Carl en français), mais d'un « mauvais récit » que se raconte la machine. L'œuvre permet donc de concevoir les risques causés par une superintelligence non en termes de domination technique mais de construction d'une identité narrative défaillante.

Genèse d'un chef-d'œuvre

En mars 1964, Stanley Kubrick décide de réaliser le film de science-fiction de référence, inspiré par la nouvelle *The Sentinel* d'Arthur Charles Clarke, publiée en 1948[2]. Les deux auteurs conviennent d'écrire ensemble un roman qu'ils transformeront ensuite en scénario – un processus d'élaboration très rare au cinéma. Une première version est soumise à la Metro-Goldwin-Mayer, qui accepte, en février 1965, de financer le film *2001, l'odyssée de l'espace*, dont le titre fait référence au récit d'Homère.

Clarke propose de conclure le film sur une régression de l'astronaute Dave Bowman jusqu'au stade prénatal. Pour cela, Bowman doit rester le dernier membre en vie à bord du vaisseau Discovery parti en mission pour Jupiter (voir encadré à la page suivante). Pour éliminer le reste de l'équipage, les auteurs inventent le personnage d'Hal 9000, superordinateur homicide dont la caméra-œil évoque le cyclope. Le tournage débute en décembre 1965. Kubrick s'entoure d'experts tels Marvin Minsky pour le conseiller, tandis que la NASA et de grandes entreprises comme IBM acceptent de prêter leurs dernières technologies. Les retards dus au perfectionnisme de Kubrick s'accumulent ; 2001 sort enfin

au cinéma en avril 1968 aux Etats-Unis et en septembre en France. Le roman est publié durant l'été 1968 et devient un succès en librairie. Autant le film présente des éléments énigmatiques, autant le livre choisit de tout expliciter et ôte au scénario tout son mystère.

La promotion du film vante son exactitude scientifique pour intéresser un public plus large que le cercle des amateurs de science-fiction[3]. Néanmoins, à sa sortie, le film divise. Il surprend notamment par sa lenteur, interroge les spectateurs sur le sens de la séquence finale. Après des débuts difficiles, le film est finalement bénéficiaire – il a coûté 12 millions de dollars et rapporte 40 millions[4]. Nominé pour quatre oscars, il n'est récompensé que pour ses effets spéciaux. Mais au fil des années, il s'impose comme un chef-d'œuvre du cinéma. Cinquante ans après sa sortie, le 71e Festival de Cannes rend hommage au film de Kubrick en le projetant le 12 mai 2018 à partir d'une copie neuve tirée du négatif original, et en donnant carte blanche au réalisateur Christopher Nolan qui le discutera lors d'une *master class*.

Synopsis du film : A l'aube de l'humanité, des hommes-singes luttent pour leur survie, lorsqu'un étrange monolithe apparaît. L'un d'eux le touche et a ensuite l'idée d'utiliser un os comme arme pour se défendre. Des millions d'années plus tard, une équipe de chercheurs menée par Floyd Heywood se rend sur la Lune pour examiner un monolithe qui a été découvert ; l'objet émet soudainement un son strident. Dix-huit mois après, le vaisseau Discovery quitte la Terre en direction de Jupiter. L'équipage est composé de Dave Bowman, Frank Poole, trois astronautes en hibernation et Hal 9000, une superintelligence artificielle. Réputé infaillible, Hal fait malgré tout une erreur d'analyse et les deux humains décident par prudence de le déconnecter. L'ordinateur réplique en éliminant l'équipage. Bowman, le dernier survivant, parvient à retirer les blocs mémoires de Hal, qui s'éteint définitivement. Un message enregistré est alors diffusé. Heywood apparaît sur un écran et dévoile l'objectif de la mission, jusqu'ici connu seulement de Hal : un objet extraterrestre découvert sur la Lune émet un puissant signal en direction de Jupiter et c'est pour cela que le Discovery doit s'y rendre. Bowman continue seul son voyage. A proximité de Jupiter, il découvre un nouveau monolithe, avant d'être projeté dans une autre dimension spatiotemporelle. Il se retrouve dans une chambre du XVIIIe siècle. Il vieillit, meurt et renaît sous la forme d'un fœtus. Dans la dernière séquence, le fœtus flotte dans l'espace et contemple la Terre.

Intelligence artificielle et identité narrative

Comment le personnage artificiel Hal s'inscrit-il dans la narration de ce film événement ? En tant que système nerveux central du Discovery, Hal a pour fonction d'assurer le bon déroulement de la mission en veillant à la survie de l'équipage durant le voyage. Quintessence de l'IA, Hal peut reproduire plus rapidement la plupart des activités cérébrales humaines. Il a été éduqué, ce qui renvoie à la notion de « machine-enfant » proposée par Alan Turing en 1950 : « *au lieu de produire un programme qui simule l'esprit adulte », suggérait le savant britannique, « pourquoi ne pas plutôt essayer d'en produire un qui simule celui de*

l'enfant ? S'il était alors soumis à une éducation appropriée, on aboutirait au cerveau humain »[5]. Sa capacité à s'exprimer dans la même langue que les astronautes, ses fonctions cognitives, son intérêt pour les croquis de Bowman, les plans subjectifs qui lui sont réservés, renforcent l'idée que Hal est un « individu » doté d'une conscience propre et d'une épaisseur, au-delà de ses performances techniques. Tel un être humain, à plusieurs reprises, il laisse entrevoir son identité à travers la mise en récit de ses actes. En particulier, le personnage qu'il élabore en s'adressant aux humains se distingue par son infaillibilité : Hal affirme que les ordinateurs de sa série n'ont jamais commis d'erreur ou déformé une information.

La collaboration entre l'homme et la machine va cependant se rompre car Hal est soumis à un conflit de programmation. L'ordinateur a été entraîné à la fois à dire la vérité et à tromper l'équipage[6]. Contrairement aux humains Poole et Bowman, Hal connaît le véritable objectif de la mission, à savoir découvrir des signes d'intelligence extraterrestre près de Jupiter. Afin d'éviter que les deux hommes ne dévoilent cet objectif lors de leurs communications avec leur famille et les journalistes, ils n'en ont pas été informés. Mais cette dissimulation donne à Hal un sentiment d'imperfection, ce que nous interprétons comme une fissure de son « identité narrative »[7]. Le roman nous éclaire sur ce point : « *Durant les cent derniers millions de milles, il avait ruminé le secret qu'il ne pouvait partager avec Poole et Bowman. Il vivait dans le mensonge et, très bientôt, ses collègues sauraient qu'il avait aidé à les trahir. (…) Il avait seulement conscience du conflit qui, lentement, détruisait son intégrité, le conflit entre la vérité et la vérité dissimulée* »[8].

Alors qu'il évoque les incertitudes de la mission dans une conversation avec Bowman, Hal commet une erreur de pronostic. Il détecte une déficience dans l'unité AE-35 et prédit qu'elle tombera en panne dans soixante-douze heures. Poole et Bowman examinent l'unité et ne détectent aucun défaut, ce que confirme sur Terre le jumeau de Hal. Comment interpréter cette erreur ? L'unité AE-35 permet au Discovery de communiquer avec la Terre. La faute de Hal relèverait ainsi d'un passage à l'acte : couper le lien avec la Terre autoriserait l'ordinateur à révéler le secret aux deux astronautes et à résoudre le conflit qui l'anime. Hal refuse toutefois d'admettre son erreur. Piégé dans son récit, il rejette la faute sur l'humain. Mais l'« erreur humaine » semble ici avoir changé de camp…

De la violence généralisée au sacrifice de la machine

L'évidence d'une erreur de la machine, pourtant réputée infaillible, anéantit l'imaginaire de maîtrise qui présidait à la mission du Discovery. Passée la stupeur, Poole et

Bowman décident par précaution de déconnecter Hal, au risque de plonger dans l'inconnu. En dépit de leurs précautions, l'ordinateur parvient à percevoir leur intention ; la fin du récit qui se dessine le terrorise, comme l'explique Clarke : « *Pour Hal, c'était l'équivalent de la mort. Il n'avait jamais dormi et ignorait que l'on pût s'éveiller…* »[9]. C'est la voie ouverte à la violence généralisée entre l'homme et la machine. Hal élimine Poole ainsi que les trois membres de l'équipage placés en hibernation pour ce long voyage et qui devaient être réveillés lorsque le vaisseau arriverait à proximité de Jupiter. Dernier survivant, Bowman parvient in extremis à se rendre dans l'unité centrale de Hal et débranche les blocs de circuit mémoire de leur logement. Tel un enfant pris en défaut, l'ordinateur tente, par un nouveau récit de lui-même, d'infléchir le projet de Bowman de le lobotomiser : « *Je sais que ne n'ai pas toujours été irréprochable (…) Je me sens maintenant beaucoup mieux. (…) Je sais que j'ai pris de très mauvaises décisions récemment. Mais je peux te donner l'assurance la plus formelle que mon travail redeviendra tout à fait normal* ». Hal échoue et sa voix finit par s'éteindre définitivement. Bowman est certes victorieux, mais il a dû sacrifier la dernière intelligence avec laquelle il pouvait interagir et doit terminer seul son voyage.

Des récits pour penser l'IA

Production artistique, *2001* offre également la narration d'un rapport possible entre l'IA et les hommes. D'ailleurs, les risques liés à une IA forte et généralisée à tous les niveaux de la société ne peuvent être anticipés qu'à travers la création de récits. C'est à cet exercice que s'est prêté Yuval Noah Harari, en imaginant dans son dernier livre un futur où l'automatisation des machines causerait la disparition de la majorité des emplois[10]. Selon cet historien, les hommes risquent de perdre leur valeur économique car l'intelligence sera découplée de la conscience. Si l'intelligence est nécessaire pour conduire une voiture ou diagnostiquer une maladie, la conscience et les expériences subjectives restent en revanche optionnelles pour accomplir ces tâches.

D'autres auteurs avancent toutefois l'idée que les hommes n'auront pas l'occasion de connaître cette révolution aux conséquences désastreuses car ils pourraient bien être exterminés dès l'apparition d'une intelligence artificielle supérieure à l'intelligence humaine. Nick Bostrom espère que d'ici là nous aurons acquis la maîtrise suffisante pour survivre à cette explosion d'intelligence[11]. Dans son ouvrage, il précise également qu'une IA n'a ni à ressembler à l'esprit humain, ni à être motivée par des sentiments humains – qui demanderaient des efforts contreproductifs pour être réalisés.

Yuval Harari et Nick Bostrom fondent leurs conclusions sur une réduction de l'action à sa dimension fonctionnelle, jugée sur le seul plan de l'efficacité. Une telle vision délaisse des pans entiers de l'existence. Pour le neuroscientifique Antonio Damasio, ces fictions sont ainsi très discutables d'un point de vue scientifique. Selon lui, « *les actes de l'intellect doivent être liés de manière fonctionnelle à ceux de la perception des émotions, sans quoi il est impossible d'aboutir à un fonctionnement ressemblant de près ou de loin à celui d'un organisme vivant (et notamment à celui d'un être humain)* »[12], c'est-à-dire intégrant une forme de subjectivité. Or, « *sans la subjectivité, rien n'a d'importance ; en l'absence d'un certain degré d'expérience intégrée, nous sommes incapables de réflexion et de discernement – et donc de créativité* »[13]. Comment dès lors l'IA pourrait-elle redéfinir sa propre mission au détriment de ses créateurs ? Le risque principal serait alors celui d'une guerre cybernétique initiée par les humains eux-mêmes.

Alternativement, le récit de *2001* envisage l'intelligence dans sa dimension narrative, y compris lorsqu'elle se fait artificielle. L'IA peut alors élaborer des récits originaux d'elle-même, mais également faillir du fait des contingences et de ses propres errements ; et l'homme est susceptible de découvrir des ressorts d'entrée en résilience pour le cas échéant l'affronter et au final la vaincre.

Nul ne peut prédire avec certitude ce que pourrait être une superintelligence artificielle. Mais il est possible de l'imaginer en produisant des récits stimulants notre réflexivité. Confronter les récits scientifiques aux fictions des romanciers ou des cinéastes paraît plus que jamais nécessaire pour nous figurer les catastrophes en devenir, et penser la place souhaitable d'une nouvelle technologie susceptible de modifier la condition humaine. Comme l'énonce le philosophe Jean-Pierre Dupuy, « *Penser ce que nous faisons, c'est aujourd'hui penser notre action technique sur le monde et sur nous-mêmes, et cela requiert une capacité narrative que les humanités aident fortement à développer* »[14].

Aurélien Portelli, chercheur associé au Centre de recherche sur les Risques et les Crises (CRC) de MINES ParisTech, **Sébastien Travadel**, maître assistant au CRC de MINES ParisTech, et **Franck Guarnieri**, directeur du CRC de MINES ParisTech.

[1] Nick Bostrom (2017). *Superintelligence*, Paris, Dunod, 495 p., p. 169.

[2] Piers Bizony (2000). *2001, le futur selon Kubrick*, Paris, Cahiers du Cinéma, 167 p.

[3] Michel Chion (2005). *Stanley Kubrick. L'humain, ni plus ni moins*, Paris, Cahiers du cinéma, 559 p.

[4] Iris Mazzacurati, « Il était une fois 2001, L'Odyssée de l'espace », *L'Express*, 25 mars 2011 [en ligne].

[5] Nick Bostrom, *op. cit.*, p. 43.

[6] Thierry Noisette, « L'Odyssée de l'espace vu par Stanley Kubrick : "Hal a un complexe de culpabilité" », *L'Obs*, 12 août 2017 [en ligne].

[7] Paul Ricœur (1990). *Soi-même comme un autre*, Paris, Editions du Seuil, 424 p.

[8] Artur C. Clarke (1991). *2001, l'odyssée de l'espace*, Paris, Editions J'ai Lu (1ère éd. 1968), 221 p., p. 148.

[9] *Ibid.*, p. 149.

[10] Yuval Noah Harari (2017). *Homo Deus*, Paris, Albin Michel, 463 p.

[11] Nick Bostrom, *op. cit.*

[12] Antonio Damasio (2017). *L'ordre étrange des choses. La vie, les sentiments et la fabrique de la culture*, Paris, Odile Jacob, 392 p., p. 287.

[13] *Ibid.*, p. 206.

[14] Jean-Pierre Dupuy (2010). *La marque du sacré*, Paris, Flammarion, 280 p., p. 63.