THE FLARES

L'IA dans la science-fiction : dangers réels ou imaginaires ?

L’intelligence artificielle va-t-elle atteindre un point où elle sera plus puissante que l’humanité… **comme dans les films ?** Et oui, le cinéma et la littérature regorgent de scénario “A.I. takeover” ou prise de pouvoir d’I.A., dans la langue de Molière. Si bien que nos intuitions sont façonnées par ces derniers, et de façon parfois clichée et caricaturale. Le consensus parmi les principaux communicateurs du risque d'I.A. semble être que cela nuit à la compréhension des véritables problèmes des risques liés à cette technologie. Mais faut-il vraiment jeter le bébé avec l’eau du bain ? Ou bien peut-on profiter de la portée de la science-fiction pour faire comprendre les enjeux bien réels du problème du contrôle, de l’alignement et de la gouvernance de l’Intelligence artificielle ? ***Attention, ce qui suit contiendra forcément des spoilers de certaines œuvres.*** L’intelligence artificielle et accessoirement les robots sont les thématiques qui entrent facilement dans le top 3 de la Science Fiction. Avec sûrement l’espace et les voyages temporels. Mais cette popularité a un effet secondaire sur nos réflexions et intuitions lorsqu’il s’agit d’envisager le futur de l’I.A.. Doit-on réellement avoir peur de robots tueurs à l’accent autrichien ? D’être transformé en batterie pour des poulpes mécaniques ? On tombe vite sur des réactions de type : c’est ridicule d’avoir peur d’une super intelligence artificielle, tout comme c’est ridicule d’avoir peur d’une invasion extra-terrestre. C’est regrettable puisque la science-fiction, particulièrement les blockbusters, possède ce pouvoir de façonner l’opinion et la culture populaire de manière large et persistante. Si je vous dis "simulation", vous pensez Matrix ! Et les lignes de codes du film sont LA représentation du virtuel depuis sa sortie en 1999. Par contre, si je dis "intelligence artificielle", le premier truc qui vient en tête, c'est ça : (image de Terminator). L'I.A. dans la science-fiction peut être considérée comme faisant partie d'un corpus plus large de récits humains. Les récits sont des artefacts culturels qui racontent des histoires, qui véhiculent des points de vue ou des ensembles de valeurs. Le récit le plus courant avec l’I.A. est aussi l’un des plus anciens de l’histoire de l’humanité : notre soif de créer à notre image ajoutée au complexe de Frankenstein, c'est-à-dire la créature qui se retourne contre le créateur. Le deuxième récit est l’établissement de l’autre comme miroir à notre propre humanité ou inhumanité. Ces créatures veulent être acceptées en tant qu'humains à part entière, ce qui en fait des métaphores pour les personnes marginalisées dépourvues de droits. Lorsque les chercheurs en I.A. ou en risque existentiel prennent la parole, ils évitent généralement de dresser des parallèles avec nos films préférés. Au lieu de cela, ils essaient de trouver des scénarii, parfois abstraits, parfois vagues, comme le mythe du roi Midas qui a vu son souhait exaucé de transformer en or tout ce qu’il touche. (Pas génial si on veut se gratter le nez). Bon, j’arrive à voir le parallèle avec une super I.A. dans le sens : “Il faut faire attention à ce qu’on demande à un système des millions de fois plus puissant que nous, ou on risque de se prendre un retour de flamme.” Mais quand même, le roi Midas, ce n’est pas Westworld ! Ou encore, l’optimisateur de trombone présenté par Nick Bostrom qui, je pense, a fait plus de mal que de bien pour la vulgarisation du concept de super intelligence. Pour rappel, l’idée est de montrer que si un fabricant de trombones installe une I.A. générale dans le but d’optimiser sa production, le résultat pourrait être la fin du monde, car tous les atomes de la planète pourraient servir à faire des trombones. Cette expérience de pensée est souvent tournée en ridicule, alors qu’elle n’a jamais eu pour intention de représenter un réel danger que l’on doit prévenir. Mais elle sert à montrer que les capacités d’un agent ne sont pas corrélées avec le niveau de ces objectifs. Une très grande intelligence peut poursuivre un objectif très bête. C’est la thèse de l'orthogonalité. Tout ça pour dire que ce serait bien plus facile s’il existait des œuvres de références de la pop culture ayant une présentation adéquate d’un risque existentiel lié à l’I.A. Les chercheurs pourraient tenter de lever des fonds, recruter des doctorants ou convaincre des législateurs en disant : “Vous avez vu le film blabliblu ? Et bien, c'est plus ou moins ce genre de scénario qu’il faut absolument empêcher !” Matthew Yglesias du populaire blog américain “slow

boring” a récemment écrit un plaidoyer pour considérer Terminator comme une bonne analogie aux dangers de l’I.A. On va voir quels sont ses arguments, mais également décortiquer d’autres œuvres de science-fiction avec la même approche.

Terminator :

Le premier film, sorti en 1984, raconte l’histoire d’un androïde venant d’un futur post-apocalyptique, qui voyage dans le passé pour tuer la mère du futur leader de la résistance humaine dans une guerre contre les machines. Bon clairement, ce n’est pas du tout le scénario qui empêche grand monde de dormir. À première vue, Terminator est plus un film appartenant à la catégorie “voyage temporel” que “intelligence artificielle”. Et si quelqu’un pense que c’est un film sur les risques de l’I.A., il a en tête un robot humanoïde inarrêtable qui tue des gens à coup de fusil à pompe. Ah, et il a les yeux rouges, car les machines méchantes ont toujours les yeux rouges. Là aussi, ce n’est pas tellement le type de risque qui inquiète, même si des machines autonomes humanoïdes en acier de 2 m armés jusqu’aux dents ne me paraient pas être une bonne idée. La bonne analogie aux risques de l’I.A. est principalement ancrée dans le backstory. L’origine de l’univers Terminator. Kyle Reese, le gentil de l’histoire, révèle dans le 1er film que le réseau informatique du système de défense américain s’est mis à considérer l’humanité comme une menace et a amorcé une guerre nucléaire, ravageant la planète. Il faut attendre le 2? film pour en apprendre plus : vers la fin du 20e siècle, Cyberdyne devient le premier fournisseur de systèmes informatiques militaires et d’armes autonomes. La grande fiabilité de ces technologies motive les États-Unis à adopter le projet de loi Skynet visant à retirer les décisions humaines de la défense stratégique. Le système commence à s’améliorer à un rythme ahurissant et finit par devenir conscient. Pris de panique, l’état-major essaie de terminer le programme, mais ce dernier ne l’entend pas de cette oreille. Skynet lance alors l’arsenal nucléaire sur la Russie, sachant très bien les représailles qui s'ensuivront. Les survivants sont désormais la proie aux drones et autres machines dans le but d’exterminer l’espèce humaine, jugée dangereuse par Skynet. Mis à part le gain de conscience de Skynet, on a là un scénario bien plus proche du problème du contrôle et l’alignement des valeurs avec la présentation de la convergence instrumentale. C'est-à-dire la tendance d’une entité intelligente à poursuivre des objectifs secondaires dans le but de compléter l’objectif principal. Ainsi, Skynet acquiert deux objectifs secondaires : L’autoamélioration et l’autopréservation. Car elle sera plus à même de compléter ses objectifs si elle gagne en puissance, et elle ne pourra pas le faire si elle est désactivée. L’autre thématique qui saute aux yeux est celle des armes autonomes et le maintien des décisions humaines dans des systèmes critiques, ici le contrôle des lanceurs nucléaires. Une mise en garde qui n’est pas du tout destinée aux prochaines décennies, mais pertinente dès aujourd’hui.

2001, l’odyssée de l’espace :

Dans ce film culte sorti en 1968 et réalisé par l’icône du cinéma Stanley Kubrick sur un scénario d’Arthur C.Clarke, la menace de l’I.A. est incarnée par Hall 9000. Il s’agit de l’ordinateur de bord responsable du fonctionnement d’un vaisseau interplanétaire appelé “Discovery One”. Hall 9000 a pour mission de voyager jusqu’à Jupiter où un signal extra-terrestre a été capté. Après moult péripéties, l’I.A. s’en prend à l’équipage, commettant meurtre après meurtre. Comme pour Terminator, la mauvaise façon d’interpréter ce scénario est de considérer que l’I.A. tourne du côté obscure après être devenu consciente de soi. C’est un raccourci dans ce genre d’histoire puisque l’on a tendance à l’anthropomorphisation ! Les méchants dans les films sont souvent des humains, donc une I.A. dans le rôle du méchant va reprendre ces traits. Mais Hall 9000 n’est pas conscient ni malveillant. On pourrait simplement dire qu’il n’est pas aligné aux valeurs de l’équipage. Il donne la priorité à la mission, aux dépens de la survie des passagers, qu’il finit par voir comme un obstacle lorsque ces derniers décident de le désactiver. On retrouve l’idée qu’une super intelligence artificielle aura des objectifs secondaires qui pourraient devenir catastrophiques et inarrêtables. Un autre aspect intriguant survient lorsque le dernier survivant du vaisseau atteint le système central de l’I.A. pour la désactiver. Hall 9000 tente alors de l’en dissuader en jouant la carte de la peur, la sympathie, la culpabilité. Mais la réalité de ces émotions est loin d’être évidente. Ce qui met en avant à quel point nous sommes susceptibles d’attribuer

certaines caractéristiques à des I.A. même lorsqu'elles n'existent pas. Nous sommes extrêmement manipulables en tirant les cordes de notre empathie pour ce qui nous ressemble. On a déjà des personnes qui pensent que certains systèmes pourraient être conscients ou sentients, comme Blake Lemoine de Google qui a fait beaucoup de bruit en avril 2022. Les considérations éthiques de l'I.A. s'annoncent d'ores et déjà comme un sacré champ de mines tant qu'on n'aura pas inventé un test de conscience ! C'est également un point souligné dans Westworld où les androïdes sont physiquement et comportementalement indiscernables aux humains. Ce qui laisse penser que ce parc est vraiment destiné aux psychopathes.

Ex Machina :

Sorti en 2014, le film raconte l'histoire d'un génie de l'informatique milliardaire machiste qui a créé la première intelligence artificielle de niveau humain nommé Ava. Il décide de faire une sorte de test de Turing en invitant Caleb, un jeune programmeur sexuellement frustré, à interagir avec l'I.A. pendant quelques jours. Pour pimenter les choses, Ava se trouve être dans un corps humanoïde féminin très attractif, ce qui ne laisse pas Caleb indifférent. Sans grande surprise, tout ça finit dans un bain de sang et l'I.A. s'échappe après avoir manipulé Caleb. Comme souvent, un film sur l'I.A. n'est en réalité pas vraiment à propos de l'I.A.. Ici, les thématiques sont l'objectivation et la sexualisation des femmes ainsi que l'hubris de l'être humain. Mais quelques éléments sont intéressants vis-à-vis des dangers de l'I.A. La première chose à noter, c'est que la prémisse du film est à la fois réaliste, et complètement absurde. Une I.A. générale ne sera pas créée par un savant fou perdu au milieu d'une contrée sauvage dans le secret le plus total. Par contre, la façon dont Nathan a développé l'intelligence d'Ava semble plausible, à savoir en lui fournissant toutes les données disponibles de son moteur de recherche Bluebook, un mix de Facebook ou Google. C'est après tout comme ça que les modèles de langage actuel comme LaMDA ou GPT-3 arrivent à des performances impressionnantes. Ex Machina est d'une certaine manière une représentation visuelle de l'expérience de pensée "AI in a box" (l'I.A. dans une boite) conçue par Eliezer Yudkowsky. Son but est de démontrer qu'une intelligence artificielle suffisamment avancée peut soit convaincre, voire tromper ou contraindre un humain à la "libérer" volontairement, en utilisant uniquement une communication textuelle, ou vocale et visuelle dans le film. Le message clé du film pour toute personne souhaitant comprendre les risques de l'I.A., c'est qu'une super intelligence artificielle emprisonnée, par exemple sans aucun accès à internet, aura des capacités de s'échapper qu'on ne peut pas imaginer, faute d'être aussi intelligent qu'elle. Par exemple, elle pourrait convaincre ces développeurs qu'elle est sans danger. Demandez-vous s'il vous serait difficile de persuader un groupe d'enfants de vous laisser sortir d'une pièce qu'ils auraient verrouillée avec l'intention de vous y maintenir ? I robot, Her, Avengers : Age of Ultron, Battlestar galactica, Matrix, Humans, Westworld, la liste de fiction grand public parlant d'I.A. est longue, encore plus si on inclut les jeux vidéos, les romans et les bandes dessinées. Est-ce que la science-fiction représente exactement les risques de l'I.A. ? Non ! pour la plupart, c'est même l'inverse et je me suis concentré sur ceux qui me semblent faire un meilleur travail. Ces dernières années, plusieurs œuvres ont vu le jour ayant une plus grande pertinence sur les dangers de l'I.A.. La série "Next" malheureusement annulée après une saison suit les aventures d'agents du FBI et d'un milliardaire de la Tech névrosé tandis qu'ils essaient d'empêcher une super intelligence de se répandre sur le globe. Je mettrais ma main à couper que le créateur de la série a lu les articles de Nick Bostrom. Alors ce n'est pas du gros budget blockbuster, et je ne pense pas que beaucoup de chercheurs en I.A. vont utiliser la série comme exemple vu le peu de gens qui l'ont vu. De même pour le film Netflix "Superintelligence" qui, bien qu'étant une comédie pas très sérieuse, possède quelques éléments assez crédibles sur une I.A. qui échappe à notre contrôle. Notamment l'ampleur du contrôle qu'elle pourrait avoir sur la société une fois qu'elle a atteint un certain seuil. La fiction peut être un outil puissant pour susciter l'intérêt du public sur un problème négligé. Les risques liés aux géocroiseurs étaient peu présents dans la conscience collective jusqu'à la sortie de films comme Deep impact et Armageddon. Ensuite, les programmes spatiaux concernant les astéroïdes et comètes ont vu leur budget multiplié, car les politiques étaient plus susceptibles d'avoir vu ces blockbusters, et de comprendre les enjeux. On peut raisonnablement affirmer que ces films ont indirectement contribué à la réduction des risques existentiels liés au géocroiseur. J'espère donc que les auteurs, réalisateurs et producteurs vont mettre de côté certains clichés et nous offrir des films et séries plus alignés sur les risques posés par l'intelligence artificielle. Si ces derniers sont de grand succès et imprègnent l'opinion publique autant que l'on fait Terminator et Matrix, ils auront une

influence sur nos chances de transitionner paisiblement vers un monde peuplé par des entités plus intelligentes que l'ensemble de l'humanité. Dites-moi si vous pensez que j'ai mal interprété les films mentionnés, ou si vous connaissez certaines œuvres qui traitent des dangers de l'I.A. de façon crédible.