L'histoire de la (super) intelligence et la question de l'éthique des machines : un cocktail autocatalyseur ?

Vincent Guérin

Publié *in* Marianne Celka et Fabio La Rocca (dir.), *Transmutations, Esprit Critique, Revue internationale de sociologie et sciences sociales,* vol. 24, n° 1, été 2016, p. 43-57 [en ligne].

Bio :
Vincent Guérin est docteur en histoire contemporaine. Chargé d'enseignement à l'université d'Angers, l'université Catholique de l'Ouest et l'ESSCA *School of management*.
Email : guerinv3m@gmail.com

Résumé :
Ce texte a pour objet d'analyser, chez les transhumanistes, le couplage de l'éthique des machines avec les risques inhérents à la *superintelligence*. La première favorisant l'émergence de la seconde. Par ce biais, nous observons une accentuation du rapprochement de l'homme et de la machine, initié par le paradigme informationnel ; un renversement même avec une machine considérée comme « smarter than us ».

Mots-clés : superintelligence, transhumanisme, éthique des machines, risque, Nick Boström.

Abstract :
This text aims to analyze, among transhumanists, coupling ethics machines with risks of superintelligence. The first encouraging the emergence of the second. By this way, we observe an accentuation of the combination of man and machine, initiated by the informational paradigm ; a reversal even with a machine considered « smarter than us »

Key-words : superintelligence, transhumanism, ethics of artificial intelligence, risk existential, Nick Boström.

« *Pourquoi avons-nous besoin d'une intelligence artificielle amicale ?*
Les humains ne seront pas toujours les agents les plus intelligents sur Terre.
Ceux qui dirigent le futur.
Que nous arrivera-t-il lorsque nous ne jouerons plus ce rôle
et comment pouvons nous nous préparer à cette transition ? »
Luke Muehlhauser et Nick Boström, 2014.

Introduction

En 2014, l'informaticien et cofondateur de Skype Jaan Tallinn a créé *The Future of Life Institute* (FLI) avec entre autres les cosmologistes Anthony Aguirre (Université de Californie) et Max Tegmark (MIT). Dans le comité scientifique se trouve une constellation de personnalités célèbres comme Stephen Hawking, des auteurs à succès comme Erik Brynjolfsson (*MIT Center for Digital Business*), mais aussi l'acteur Morgan Freeman (film *Transcendance* de Wally Pfister, 2015) et l'inventeur et chef d'entreprise Elon Musk. Jaan Tallinn était déjà à l'initiative du *Centre For The Study Of Existential Risk* (CSER) ou *Terminator studies* en 2012 à l'Université de Cambridge avec le cosmologiste Martin Rees [1]. Ces deux institutions ont pour ambition, entre autres, d'anticiper les risques majeurs qui

menacent l'humanité, notamment ceux inhérents à l'intelligence artificielle (IA). Dernièrement, Bill Gates, fondateur de Microsoft, lui-même, se dit préoccupé par l'IA. Ces deux institutions et Bill Gates ont un dénominateur commun : Nick Boström. L'auteur de *Superintelligence, Paths, Dangers, Strategies* (2014), qui a impressionné Bill Gates, est membre du comité scientifique de la FLE et du CSER. Il est professeur à la faculté de philosophie de la prestigieuse Université d'Oxford et fondateur de la *Future of humanity Institute* (FHI) qui a pour objet d'anticiper les risques majeurs qui menacent l'humanité (*existential risks*). Ses recherches portent sur l'augmentation de l'homme, le transhumanisme, les risques anthropiques et spécifiquement celui de la superintelligence. En 2008, il a codirigé avec Milan M. Ćirković *Global Catastrophic Risks* (Boström, Ćirković, 2008). Cet ouvrage dénombre dix risques catastrophiques au sens d'un bouleversement radical qui menacerait l'humanité (anthropiques ou non)[1]. Parmi les risques anthropiques recensés, Eliezer S. Yudkowsky (1979-), chercheur au *Machine Intelligence Research Institute* à Berkeley (MIRI) [2], développe le chapitre sur l'IA (Yudkowsky, 2008).

Nick Boström[3] et Eliezer Yudkowsky sont transhumanistes, un courant de pensée qui conçoit l'humain, l'humanité comme imparfaits et prône une prise en main de leur évolution par la technologie. En 1998, Nick Boström a fondé avec David Pearce la *World Transhumanist Association* (WTA) et l'*Institute for Ethics & Emerging Technologies* (IEET) avec James Hughes.

Plusieurs objectifs irriguent le transhumanisme, dont le devenir postbiologique (posthumain), la superintelligence et l'amortalité (une immortalité relative). Parmi les NBIC, deux technologies ont leur faveur. La première, la nanotechnologie (une construction à partir du bas à l'échelle du nanomètre soit un milliardième de mètre) est en devenir, et la seconde, l'intelligence artificielle générale (IAG) reste un fantasme. Nick Boström et Eliezer Yudkowsky pensent que l'IA favorisera la nanotechnologie, elle-même porteuse d'inquiétude (Drexler, 1986). Eric Drexler, transhumaniste et membre du FHI, a créé en 1986, le *Foresight Institute* afin de prévenir les risques technologiques et favoriser un usage bénéfique de la nanotechnologie. Qu'est-ce-que la (super) intelligence artificielle ? Quelles sont les corrélations entre le transhumanisme et cette inquiétude montante vis-à-vis de l'IA, ou plus exactement la superintelligence ? Comment et quand pourrait-elle émerger ? Comment s'articule le complexe dit de Frankenstein et l'éthique des machines ?

La (super) intelligence artificielle

Dans son livre portant sur la quête de l'intelligence artificielle, Nils J. Nilsson, un de ses pères fondateurs, la définit comme l'activité qui consiste à rendre les machines intelligentes, au sens d'une qualité permettant à une entité d'agir de façon appropriée et anticipée dans son environnement (Nilsson, 2009).

En 1943, Warren S. McCulloch et Walter Pitts posent les premiers jalons théoriques de cette nouvelle discipline en proposant une représentation mathématique et informatique d'un neurone biologique, un neurone artificiel (formel), mais aussi sa mise en réseau en vue d'un apprentissage (Russel & Norvig, 2006 : 19).

C'est lors d'un séminaire à l'université de Dartmouth (ÉU), en 1955, que l'expression d'intelligence artificielle est forgée par John McCarthy comme une machine utilisant un

langage, étant capable de résoudre des problèmes et auto-apprenante (McCarthy et al, 1955). La jugeant inappropriée, Stuart Russel et Peter Norvig substituent à ce syntagme rationalité computationnelle (Russel & Norvig, 2006 : 20).

Si objectivement les progrès réalisés dans ce champ sont extraordinaires, le fantasme qui conduit l'attelage et génère les financements n'a pas été atteint.

Le philosophe John Searle dissocie au sein de l'IA la « faible » (*weak*) et la « forte » (*strong*). La première est omniprésente, elle s'incarne notamment dans les prouesses de deux projets développés par IBM que sont *Deep Blue* qui remporte, dans un partie controversée, un tournoi d'échecs sur le champion Garry Kasparov en 1997, et Watson qui gagne l'équivalent de « Questions pour un champion » aux États-Unis en 2011, mais aussi le défi DARPA (*Defense Advanced Research Projects Agency*) du véhicule terrestre autonome dans un environnement désertique remporté en 2005 par Stanley, l'ancêtre de Google Car. Dernièrement, le conseil d'administration d'une entreprise sud-coréenne s'est doté d'un algorithme nommé VITAL (2014). Celui-ci « siège » au conseil d'administration et dispose d'une voix au même titre que les autres administrateurs. La seconde, l'IA « forte » ou intelligence artificielle générale (IAG) ou super-ultraintelligence reste spéculative.

Le statisticien Irving J. Good serait le premier à avoir formulé en 1965 l'idée d'une ultra-intelligence mais aussi le concept d'« explosion de l'intelligence » autrement appelée singularité technologique (Good, 1965 p. 33). Le philosophe David Chalmers, analyste non transhumaniste de la singularité, attire notre attention sur le fait que l'auteur, bien que reconnu académiquement à ce moment, rédige un article largement incompris (Chalmers, 2010, p. 3).
Nick Boström, dans une focale plus large, définit la superintelligence comme tout intellect qui excède radicalement les capacités cognitives humaines et ce dans tous les domaines (Boström, 2014, p. 22).

La superintelligence pourrait advenir de différentes manières : digitalisation du cerveau et téléchargement de l'esprit (*mind uploading*) – ce qui suppose un dépassement du dualisme cartésien, et l'idée que le cerveau humain est une machine – augmentation cognitive par l'accélération de la sélection génétique, l'interconnexion *via* le réseau (l'intelligence collective), les interfaces hommes-machines (implants) et l'IA. Parmi les différentes formes possibles, la digitalisation du cerveau et l'IA seraient les voies la plus prometteuses. L'IA serait, selon Nick Boström, la plus rapide à être mise en œuvre (Boström, 2014). La première conférence sur l'intelligence artificielle générale (AGI-08) a eu lieu à l'université de Memphis en 2008. Ben Goertzel, chercheur et transhumaniste, évoque quatre raisons à ce retour en grâce de l'IAG : la puissance de calcul des ordinateurs et leur mise en réseau, le développement des sciences cognitives et neurosciences qui permettent de mieux connaître le cerveau, celui des mondes virtuels, la robotique et le développement algorithmique (Garis & Goertzel, 2009).

L'explosion de l'intelligence peut être schématisée ainsi : lorsque la machine aura atteint le stade de l'intelligence humaine (*Human-Level Machine Intelligence*) elle pourrait se reprogrammer, puis à nouveau se reprogrammer à partir de cette nouvelle programmation (AI+) et ainsi de suite jusqu'à atteindre une singularité (AI++) et ainsi produire une évolution exponentielle, voire superexponentielle.

L'idée de singularité a été popularisée par le mathématicien et auteur de science fiction Vernor Vinge en 1993 et l'inventeur et futurologue Ray Kurzweil en 2005. Un sondage réalisé par

Nick Boström auprès de chercheurs de l'IA fait apparaître que 10 % d'entre eux pensent que ce stade pourrait être atteint en 2030, 50 % en 2050 et 90 % en 2100. Eliezer Yudkowsky utilise l'expression de boucle d'auto-amélioration (*recursive self improvement*) pour décrire ce phénomène : une IA qui réécrit son propre « algorithme cognitif » (Yudkowsky, 2011). En 2008, l'économiste Robin Hanson (FHI) et Eliezer Yudkowsky ont longuement débattu à ce propos, l'un considérant que le décollage sera lent, qu'il prendra des années, des décennies et l'autre, au contraire, fulgurant (Yudkowsky & Hanson, 2013). Boström, quant à lui, pense aussi que le décollage sera explosif : minutes, heures, journées (Boström, 2014, p. 64-65).

Si les estimations de cette explosion de l'intelligence varient, elles sont cependant toutes à court terme. Irving Good (1965) prédisait l'avènement de l'ultraintelligence autour de l'année 2000, Vernor Vinge (1993) entre 2005 et 2030, Yudkowsky 2021 et Ray Kurzweil (2005) vers 2030 (Chalmers, 2010, p. 6). David Chalmers tempère l'intensité de la singularité en posant, en contrepoint, les limites de la physique (Chalmers, 2010, p. 3.). Il va sans dire que ces idées sont loin de faire l'unanimité dans le champ de l'IA (ex. McDermott, 2006)

© Sébastien Genest, 2015.

Le complexe dit de Frankenstein et l'éthique des machines

L'IAG est pensée par certains transhumanistes, tel Peter Diamandis, comme le facteur qui va favoriser les technologies de rupture (NBC). Le cofondateur de l'Université de la singularité, profondément optimiste, prédit l'abondance à venir : extension de nos capacités, accès à de nouvelles ressources (Diamandis, 2012, 2015). Plus rien ne sera impossible. L'IAG apparaît non comme une menace mais l'opportunité d'une collaboration, puis d'une co-évolution qui "nous permettra d'être plus humain car plus éthique et plus moral" (Diamandis, 2015).
Boström et Yudkowsky sont plus ambivalents. Bien que considérant l'IAG comme l'événement le plus important de l'histoire, sans négliger les aspects positifs, ils considèrent aussi le côté potentiellement eschatologique. Selon eux, il faut s'attendre au développement d'une IA potentiellement hostile (*unfriendly).* Déjà en 2000, Bill Joy, inventeur du système

Java, après avoir entendu Ray Kurzweil, s'interrogeait : "Pourquoi le futur n'a pas besoin de nous" (Joy, 2000).

C'est l'irruption du complexe dit de "Frankenstein", forgé par l'auteur de science fiction Isaac Asimov. Celui-ci est inspiré du roman *Frankenstein ou le Prométhée moderne* (1818) de Mary Shelley qui met en scène une créature se retournant contre son créateur, le punissant ainsi de l'impudence de s'être hissé à la hauteur de Dieu en créant la vie.

En 1921, le dramaturge tchèque Karel Capek dans sa pièce *Rossum's Universel Robots*, qui donne naissance au terme de robot (en tchèque : corvée, servitude), met à nouveau en scène ce ressort dramatique. Le jeune Asimov, considérant les machines uniquement comme telles et las de la récurrence *ab nauseam* de ce scénario, décide de faire des machines rassurantes en utilisant l'artifice des lois de la *robotique* – trois lois plus une loi 0 – (*Runaround*, 1942). Ce sera le fil conducteur de trente cinq nouvelles et cinq romans sur les robots. Asimov considérait ce concept comme le plus influent de son œuvre (Asimov, 1990). Seulement, les lois de la robotique qui avaient pour objet de réduire le fantasme, l'ont au contraire grandement alimenté. Jugées inopérantes dans le réel, elles ne sont pas prises au sérieux dans le champ de l'IA, l'abstraction qu'elles induisent rend difficile leur implémentation dans un programme (McCauley, 2007 : 10).

Malgré tout, les lois d'Asimov, qui viennent contrecarrer dans la fiction la prise de pourvoir par les machines, inspirent l'éthique des machines qui émergera dans les années 2000. Jusqu'alors la relation entre l'éthique et la technologie portait sur la responsabilité ou l'irresponsabilité humaine, certains, plus aventureux, interrogeaient la manière de traiter les machines. Dans les deux cas, seul l'humain était engagé dans une démarche éthique. Dans une perspective troublante, l'éthique des machines vise à créer des agents intelligents qui auront un comportement « moral » vis-à-vis des humains et des autres machines (Anderson, Anderson & Armen, 2004).

Une des premières occurrences apparaît lors d'une conférence de Warren S. McCulloch en 1952 (McCulloch, 1952). En 2000, Storrs Hall, qui a été président du *Foresight Institute*, forge l'expression *Machine Ethics* (Anderson & Anderson, 2010, p. 77). C'est véritablement en 2004 que la philosophe Susan Leigh Anderson et l'informaticien Michael Anderson lancent cette recherche (Anderson & Anderson, 2004). L'année suivante se tient un premier congrès (Anderson & Anderson, 2011). En 2006, dans un numéro spécial de la revue *IEEE intelligence systems*, deux articles portent sur l'éthique des machines dont un signé par Colin Allen, Wendell Wallach et Iva Smit *Why machine ethics ?*. Wendell Wallach, éthicien, est membre de l'IEET. En 2009, Colin Allen et Wendell Wallach publient le premier livre sur le sujet : *Moral Machines*. En 2010, [le robot humanoïde] Nao a été le premier a bénéficié d'un "comportement" guidé par des principes éthiques dans le cadre d'une gestion de prise de médicaments chez des patients (Anderson & Anderson, 2010).

Sur le terrain éminemment humain de la morale, de l'éthique, on assiste à un nouveau rapprochement symbolique de l'homme et de la machine, initié par le paradigme informationnel. Lors de l'AGI-08, le roboticien militaire Ronald C. Arkin soutient que les robots seront plus humains aux combats que les humains eux-mêmes (Arkin, 2008). En 2014, l'*Office of Naval Research*, chargé des programmes de recherche militaires, déclarait promouvoir une exploration de "raisonnements éthiques" à destination des systèmes robotiques autonomes armés. Une dotation de 7,5 millions de dollars, devant stimuler cette recherche, a été allouée à cinq universités (Tucker, 2014).

Le contrôle de l'IA alimente des spéculations philosophiques transhumanistes. Ici encore, les analyses les plus fines nous viennent de Nick Boström, Eliezer Yudkowsky, mais aussi de David Chalmers. Eliezer Yudkowsky met en garde contre l'anthropomorphisation, souvent inconsciente, que nous projetons : l'IA est un espace des possibles infiniment plus divers que l'*homo sapiens*, aussi prédire serait une gageure. Il propose le terme de *Friendly AI* pour discuter l'avènement d'une hypothétique IA générale et sa compatibilité avec les valeurs humaines, l'humanité. Selon lui, il faut anticiper le défi de l'IA "amicale". Eliezer Yudkowsky reconnaît être stimulé par des articles qui ont en commun Isaac Asimov et les trois lois de la robotique (Yudkowsky, 2008).

La première précaution indispensable est d'avoir en permanence accès au code source. Avoir la possibilité de briser une auto-modification. Limiter sa croissance, même après la phase critique de l'autoréplication. L'enjeu est de maintenir son orientation. Seulement un réseau neuronal artificiel ou algorithme génétique (évolutionniste), contrairement à un réseau arborescent ou bayésien, est difficile à percer. Pour nos deux auteurs, transparence, mais aussi prévisibilité, incorruptibilité seraient quelques uns des éléments préliminaires à l'éthique des machines (Boström, Yudkowsky, 2011).

Le scénario du confinement de l'IA, sans possibilité d'interagir physiquement avec son environnement, en contrôlant son accès à l'information, dans un monde virtuel, semble voué à l'échec (Chalmers, 2010). Malgré tout, réduire l'IA à un oracle qui répond aux questions apparaît comme la meilleure option. Une connaissance approfondie des motivations de l'IA sera nécessaire (Boström, Sanders, Armstrong, 2012) David Chalmers spécule que le contrôle de l'IA dépendra de plusieurs facteurs. Si l'IA a une base humaine (digitalisation du cerveau), elle peut partager des valeurs semblables, ce qui est plus hypothétique si la machine a « appris » d'elle même (algorithme évolutionniste). Cela dépendra aussi de l'implémentation : soit elle est directe c'est à dire programmée dans une vue utilitaire sans possibilité intrinsèque, soit elle est le fruit d'un auto-apprentissage, une évolution propre. Dans le deuxième cas, il sera impossible de prédire, ni d'avoir accès à son « comportement ». Autre piste de réflexion : la corrélation entre la rationalité et la vertu (Chalmers, 2010). Chalmers entrevoit quatre scénarios pour l'humanité dans le monde de l'après singularité : l'extinction, l'isolation, l'infériorité, l'intégration. La meilleure option apparaît être l'intégration, le devenir superintelligent par l'hybridation. En d'autres termes, la nécessité de la digitalisation du cerveau, puis son *uploading* (Chalmers, 2010, p. 33).

L'ambition de certains transhumanistes est de faire advenir la superintelligence amicale, l'accompagnement de cette émergence par une éthique des machines débouche, selon nous, sur un raisonnement autoréférent : l'IA peut-elle nous sauver de l'IA ? (Dorrier, 2014).
Plus encore, cette posture pourrait favoriser l'autocatalysation de la superintelligence ou *a minima* une orientation dans ce sens. À la lumière de ces quelques développements, l'impact de la pensée transhumaniste, qui n'est pas uniquement philosophique, ne doit pas être négligé. Les transhumanistes doivent s'interroger sur leurs responsabilités politiques.

Épilogue

Selon Nick Boström, il ne faut pas croire que l'évolution sera toujours avantageuse pour l'humanité. Sans négliger les risques naturels (astéroïdes, super-volcans, sursaut gamma, etc.), ce sont les risques anthropiques que nous devons redouter (biologie synthétique, nanotechnologie) notamment la superintelligence.

Au prisme du transhumanisme, ces risques menacent de façon prématurée l'extinction (avant l'expansion du soleil) non pas de l'humanité au sens de l'*homo sapiens*, une incarnation transitoire, mais de l' « intelligence d'origine terrestre » et sa "maturité technologique", c'est-à-dire le maximum de ses potentialités : colonisation de l'espace, modification et augmentation du corps (posthumain), etc. (Boström, 2013). Nick Boström donne ici une extension radicale à la problématique de la finitude, en la faisant évoluer du corps/espèce à l'intelligence. "Nous [les transhumanistes] refusons de croire que nous sommes ce qu'il y a de mieux, que nous sommes une sorte d'aboutissement, une création indépassable" (Boström, 2005-2006)

Dans une démarche proactive, la seule manière d'éviter le désastre est de prendre en main l'évolution de l'humanité, et pour cela, il faudrait développer ce qu'il appelle un *singleton* (en mathématique un ensemble formé d'un seul élément), un ordre mondial, présidé par une entité unique indépendante qui aurait pour objet de résoudre les problèmes globaux (Boström, 2004, 2009). Ce *singleton* pourrait prendre plusieurs formes : un gouvernement démocratique, une dictature ou…une superintelligence amicale (Boström, 2006). Nick Boström avoue que si la création d'un *singleton* pourrait réduire certains risques, il pourrait en faire advenir d'autres comme un régime oppressif global et permanent (Boström, 2004, 2009).
Imaginons un instant, une super-intelligence dite amicale, compatible avec nos valeurs. Paradoxalement, cette super-intelligence ne sera-t-elle pas plus morale que nous ? (Boström & Muehlhauser, 2013). Et si c'était le cas…

Références bibliographiques

Asimov I. *Robot visions*, NY, Roc, 1990.
Boström N. *Superintelligence, Paths, Dangers, Strategies*, Oxford : 2014.
Brynjolfsson E. & Mc Afee A. *The Second Machine Age: Work Progress, and Prosperity in a Time of Brilliant Technologies*. New York : Norton & Compagny, 2014.
Diamandis P. & Kolter S. *Abundance. The future is Better than you think*. NY : Free Press, 2012.
Drexler K. E., *Engines of creation. The coming era of nanotechnology*. NY : Anchor Books, 1986.
Kurzweil R.The *Singularity is Near. When humans transcend biology*. NY : Penguin, 2005.
Nilsson N. J. *The Quest for Artificial intelligence. A history of ideas and Achievements*. 2009.
Russel S. et Norvig P. *Intelligence artificielle*. Paris : Pearson Éducation, 2006.
Wallach W. & Allen C. *Moral machines. Teaching Robots Right from Wrong*, Oxford, oup, 2009.

Articles :

Allen C., Wallach W. & Smit I., "Why Machine Ethic ?", IEEE, 2006, [en ligne].
Anderson M & Anderson S. L. & armen C. "Toward Machine Ethics", AAAI, 2004 [en ligne].
Anderson M & Anderson S. L., "Robot Be Good, Scientific American", October, 2010 [en ligne].
Arkin R. C. "Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture", 2008 [en ligne].

Boström N., Armstrong S. et Sandberg A. “Thinking Inside the Box : Controlling and Using an Oracle AI”, 2012 [en ligne].
Boström N. & Muelhlhauser L. “Why we need friendly AI”, Think, vol. 13, Issue 36, march, p. 41-47, 2013 [en ligne].
Boström N. & Yudkowsky E. “The ethics of artificial intelligence”, 2011 [en ligne].
Boström N. “The Future of Human Evolution”, in Charles Tanguy (ed), *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*. Palo Alto, rup, 2004, p. 339-371, 2004-2009[en ligne].
Boström N. “Entretien avec Nick Boström, le transhumaniste en chef”, Argument, Politique, société et histoire, Vol. 1, no 8, automne, 2005-2006.
Boström N. “What is a Singleton”, Linguistic and Philosophical Investigations, Vol. 5, no 2, p. 48-54, 2006 [en ligne].
Boström N. “Existential Risk Prevention as Global Priority”, Global Policy, Vol. 4, issue 1, p. 15-31, 2013 [en ligne].
Chalmers D. J. “The Singularity : A philosophical analysis”, Journal of Consciousness Studies, 17, 2010, p. 7-65, [en ligne].
Diamandis P. “AI More Like Iron Man’s JARVIS Is Coming This Next Decade…Bring It On”, 12 mai, 2015 [en ligne].
Dorrier Jason, “Can AI save us from AI ?”, Singularity HUB, 9 décembre, 2014.
Garis H. de & Goertzel B. “Report on the first conference on Artificial General Intelligence (AGI-08)”, 2009 [en ligne].
Good I. J. “Speculation Concerning the First Ultraintelligence Machine”, 1965 [en ligne].
Joy B., “Why the future doesn’t need us. Our most powerful 21st century technologies – robotics, genetic engineering, and nanotech – are threatening to make humans an endangered species”, Wired, avril, 2000. [en ligne]
McCarthy J, Minsky M. L., Rochester N. et Shannon C. E. “A proposal for the Dartmouth Summer Research Project on Artificial Intelligence “, 1955 [en ligne].
McCauley L. “The Frankenstein Complex and Asimov’s Three Laws”, AAAI, p. 9-14, 2007 [en ligne].
McCulloch W. S. “Toward some circuity of ethical robots. An observational science of the genesis of social evaluation in the mind-like behavior of artefacts” , 1952 [en ligne].
McDermott D. « Kurzweil’s argument for the success of AI », Artificial Intelligence, 170, p. 1227-1233, 2006 [en ligne].
Tucker P., “Now the military is going to build robots that have morals”, Defense One, May, 13, 2014 [en ligne].
Vinge V. “The Coming Technological Singularity: How to survive in the Post-Human Era”, 1993 [en ligne].
Yudkowsky E. & Hanson R. Yudkowsky- *Hanson AI Foom Debate*, MIRI, 2013 [en ligne].
Yudkowsky E. ‘‘Artificial intelligence as a positive and negative factor in global risk’’, in Boström Nick & Ćirković M. Milan (ed.), *Global Catastrophic Risks*, Oxford, 2008.

Notes de fin :

1. D’autres institutions se sont développées comme l’AI 100 à l’Université de Stanford qui se donne pour mission d’encadrer l’IA pour les 100 prochaines années.
2. Dernièrement *Global Challenge Fondation*, en partenariat avec FHI et Oxford Martin School, a publié sur internet une étude grand public. 15 risques sont recensés.
3. Nick Boström a notamment répondu philosophiquement à l’inquiétude du politologue américain Francis Fukuyama qui voit dans le transhumanisme la pire des idéologies (Boström Nick, Transhumanisme : “The World’s Most Dangerous Idea ?”, 2004, [en ligne]).