10ème Colloque Stella Incognita

Pouvoir(s), responsabilités et cas de conscience en science-fiction

Université de Reims Champagne-Ardenne
6-7-8 avril 2022

Campus Croix-Rouge
Amphithéâtre Recherche / bâtiment 13
57 rue Pierre Taittinger 51100 Reims

Comité d'organisation
Hervé LAGOGUEY (CIRLEP),
Jérôme GOFFETTE (EVS, UMR 5600),
Danièle ANDRÉ (CRHIA),
Marine RIGUET (CRIMEL)

Quels choix face à une IA gouvernante ?

Intervention d'Emmanuelle Lescouet le 6 avril 2022.

Bonjour à tous bonjour à toutes, merci d'accueillir cette communication montréalaise au colloque !
Je vais donc essayer de répondre a cette question pompeuse : quels choix face à une IA gouvernante ?
Bien que l'intelligence artificielle soit un enjeu du développement numérique actuel, c'est aussi une des pistes de réflexion sur le futur le plus populaire. D'ailleurs, la science-fiction s'est intéressée aux robots depuis ses débuts, et par la suite, à l'intelligence artificielle (Vint, 2021). Pour autant, les possibles réels de l'informatique poussent à présent à imaginer le développement et l'extension des réseaux de neurones synthétiques, et leurs implications dans la société. Et s'ils sont très utiles au quotidien dans les diverses disciplines des humanités numériques ou des *digital classics – je vous renvoie sur ces question au travaux du GREN notamment*, ces usages concrets ne vont pas nous retenir particulièrement ici. Notons tout de même que nous avons déjà en main, même pour des chercheur.e.s en littérature non spécialistes en programmation, des IA faibles, bien qu'assez basiques.

Cependant, les intelligences artificielles de science-fiction sont hautement plus complexes, qu'elles existent comme consciences ou comme êtres totalement indépendants. Le corpus mobilisable est en effet très vaste, et cela, sur tous les supports : qu'il s'agisse des visuals novels, comme *Neo Cab* et ses systèmes de gestion des émotions, aux intelligences de vaisseaux comme dans *Idealis* (de Christopher Paolini) ou *The Way to a Small Angry Planet* (de Becky Chambers), elles posent aussi la question des limitations et des possibles d'êtres-golems créés de toutes pièces comme dans *A Closed and Common Orbit* (également de Becky Chambers) ou dans le cycle des robots d'Asimov ; ainsi, de très nombreux enjeux éthiques pour les humains et les espèces sentientes sont soulevés.

Ceci étant dit, nous nous arrêterons aujourd'hui sur un enjeu très particulier des IA : comment ferions-nous si l'une d'elles était en charge de la gouvernance d'une société humaine ? Sans tomber dans l'angoisse d'une guerre robots-humains à la *Battlestar galactica*, j'aimerais essayer de penser avec vous une cogouvernance.
Pour cela, je vais majoritairement m'appuyer sur la série *Scythe* (*La Faucheuse*, dans son édition française) de Neal Shusterman. Dans cette série, l'humanité a réussi à dépasser la mortalité et a donc dû mettre en place un système de régulation de la population. Les scythes sont des sortes d'assassins qui mettent fin à la vie de certains individus. Le tout est encadré par des règles strictes, visant à éviter les dérives. Quant au reste de la société, elle est gérée par le Thunderhead : une IA à laquelle tous les objets connectés sont reliés – elle a des caméras partout y compris dans les logements, a accès à des systèmes audio où qu'ils se trouvent, mais aussi à des outils plus communs comme les réfrigérateurs connectés ou les téléphones intelligents. Mais bien que cette IA soit présente partout en

tout temps, elle l'est toutefois avec un certain nombre de protections de la vie privée, nous y reviendrons. Si le premier tome de la série est particulièrement centré sur les faucheuses (les scythes), sur l'apprentissage de jeunes recrues dans leur tâche de régulation et sur l'exposition de l'univers – dont le Thunderhead –, le deuxième tome, quant à lui, met en scène la chute du système : le moment de bascule où les faucheuses montrent les limites de leur pouvoir et l'impunité dont elles profitent, mais aussi les tentatives de contre-pouvoir et de résistances qui peuvent se mettre en place. Ce renversement ne sera toutefois réellement pensé qu'après la chute du système dans son entièreté dans le troisième et dernier tome : *The Toll*.
Car après la chute l'IA, c'est le cœur même de la vie quotidienne de tous et toutes qui s'est tue. Et c'est cela qui fait basculer ce monde dans son effondrement, nous le verrons, un effondrement de ses valeurs tout autant qu'une mise en avant de la difficulté humaine à se gérer et à faire fonctionner une société complexe en bonne intelligence.

Cet exemple nous accompagnera tout au long de cet exposé, pour essayer de penser l'exemple d'une IA gouvernante comme une pensée de la science-fiction numérique, notamment comme une alternative au cyberpunk, au sens notamment de Yannick Rumpala. Je commencerai par une brève mise en contexte des intelligences artificielles, pour continuer par l'agentivité dont elles peuvent faire preuve. Ces possibles menant à des prises de pouvoirs – dont la gouvernance est la plus évidente – et donc à la constitution de contre-pouvoir plus ou moins efficaces. Enfin, je reviendrai sur les parallèles avec le numérique actuel et prévisible dans un futur immédiat.

1. L'intelligence artificielle : un trope de science-fiction

Comme le dit Sherryl Vint :

> « It would be difficult to overstate the degree to which the sf imaginary has shaped notion of the robot, and later of artificial intelligence » (p.77)

L'intérêt technologique du genre est évident, autant que la fascination pour la création de conscience autre qu'humaine, et la possibilité de programmer un être sentient avec du langage inventé – ici, des langages de programmation – est particulièrement fascinante.

La question des possibles techniques est au cœur de nombreux textes, et les langages permettant de travailler à la création d'intelligence-golem sont un questionnement comme l'individuation de ces consciences ; nous parlons ici de ce moment de bascule entre intelligence artificielle utilitaire – à la manière de celles que nous avons déjà aujourd'hui – et conscience autonome, de cette balance posant la question de la limite et de la définition de l'être.

Pour preuve, l'indépendance d'une conscience artificielle propose de nombreuses intrigues de leur rapport avec les humaines et/ou les aliens — les sentients organiques. De la même manière, la limite de la prise en compte de leurs individualités par rapport à de simples programmes amène également à repenser les relations interpersonnelles dans ces univers, car du moment où les intelligences artificielles acquièrent une conscience et le pouvoir d'agir via les réseaux ou via divers objets connectés, il devient tout aussi nécessaire de prendre en compte ces actions dans la société et de les encadrer par la loi.

Bien sûr, la grande diversité des intrigues autour de ces questions nous empêche d'adresser l'intégralité de la question. Cependant, nous pouvons nous concentrer sur le cas extrême du choix de laisser à une intelligence artificielle l'intégralité de l'administration notre société.

2. L'agentivité et le pouvoir

Les intelligences artificielles ont une puissance de calcul logiquement bien plus grande que nous, humain.e.s. Or, par cette puissance et la mémoire potentiellement infinie selon les serveurs auquel iels ont accès, iels peuvent « savoir » bien plus de choses que nous et prendre des décisions bien plus éclairées. Elles sont capables de surveiller des paramètres ou des situations sans limites de temps ou de fatigue, et si leurs processeurs sont suffisamment puissants, elles peuvent démultiplier leurs points d'actions ou leur aire de surveillance.

Comme tout programme informatique, elles sont déterminées par un langage écrit par ses concepteur.ice.s. L'intelligence artificielle est donc limitée par ce que son code lui autorise ou l'oblige à faire : elle est dépendante de la flexibilité, ou non, de cette conception. La responsabilité de la conception est donc immense, car une ou plusieurs personnes se retrouvent responsables des possibles cognitifs et du développement des sentients synthétiques. Dans Scythe, les Faucheses qui ont programmé originellement le Thunderhead l'ont fait sans laisser de documentation nécessaire pour le faire ensuite dévier : ielles ont tout fait pour qu'il soit inaltérable, ne laissant derrière ielles que les raisons de sa mise en place.

D'ailleurs, les actions de ces consciences artificielles posent la question de la responsabilité de leur concepteur.ice.s. Cependant, elles apparaissent comme plus que des programmes : en sont-elles encore si elles ont des émotions et des perceptions de soi ? La question est plus complexe que cela : si elles reposent sur du code et de la programmation, celui-ci il est généralement invisible et inaccessible. En tout cas hors de la compréhension des gens. Elles deviennent alors tellement complexes qu'elles sont – comme le Thunderhead après sa chute quand il commence son introspection – inaccessibles à des compréhensions plus « bas de gamme », simplement humaines. Il ne peut qu'explorer son propre code pour trouver des patterns et des logiques expliquant

ses biais et fonctionnement. La rétroingénieurie nécessaire occupant l'intégralité ou presque de sa puissance de calcul.

Le choix d'utiliser des IA est souvent dicté par leur capacité à se rendre utiles : elles sont capables de gérer les paramètres d'un vaisseau en pleine navigation interstellaire ou de prendre en charge tous les aspects de la vie quotidienne d'un.e ou plusieurs individu.e.s. Exactement comme nous avons majoritairement adopté des ordinateurs parce qu'ils nous simplifient la plupart de nos tâches quotidiennes, sans comprendre finement la mécanique qui les sous-tend.
En effet, pour reprendre notre exemple, le Thunderhead peut devancer les besoins de chacun.e car il dispose de toutes les données nécessaires : les actions faites et les gouts depuis la naissance, les questions posées ou les souhaits formulés en lignes.
Ainsi, cette IA-ci suit tous les membres de la société humaine (bien que le propos soit basé aux États-Unis) et il devient bien plus simple de laisser le Thunderhead expliquer aux personnages comment s'organiser, qu'est-ce qu'ielles veulent pour le diner ou quelle serait la meilleure formation, prenant peu à peu en charge tous les aspects décisionnels du quotidien, remplaçant même les présidents des États ou les chambres parlementaires. Il est clairement expliqué, au début du premier volume et dès l'exposition de l'univers, que cette intelligence artificielle est bien plus fiable pour gouverner parce qu'elle connait finement les envies et les réalités de chacun, chacune, et des divers groupes, sans avoir à passer par des représentations biaisées. Aussi, parce qu'elle a accès à la réalité, y compris celle tenue secrète des individu.e.s. Il est même trop prévenant : plaçant toute la population dans un presque sommeil : personne n'a plus aucun choix à faire, même les choix futiles du quotidien sont aseptisés.

Le plus souvent, la simplicité d'usage et la très grande faillibilité sont les arguments mis en avant dans l'usage et la décision de mettre en place de tels systèmes, et c'est déjà le cas pour certaines utilisations actuelles.
Exactement comme lorsque nous préférons laisser des algorithmes parcourir des corpus trop volumineux ou gérer des données trop lourdes pour être manipulées dans une vie humaine. Or, c'est en ligne que nous révélons toutes nos données : ce que nous cherchons, et donc ce que nous apprenons, ce que nous achetons, et où nous allons, où nous sommes inscrites, etc. Si ce portrait peut sembler inquiétant – et il l'est dans beaucoup de récits de science-fiction –, il est malgré tout possible d'imaginer une co-habitation paisible, avec des intelligences artificielles dont le code serait éthique, et ouvert : nous y arrivons. Il y a bien sûr et surtout des contre-pouvoirs à imaginer.

3. De possibles contre-pouvoirs.

Si dans la société de *Scythe*, le contre-pouvoir est un contre-pouvoir humain : les faucheuses n'ayant pas accès au Thunderhead, mais uniquement à ses bases de données ; d'autres contre-pouvoirs se mettent en place.

Ainsi, le partage de la gouvernance peut s'organiser avec d'autres groupes humains ou avec d'autres intelligences artificielles. Toutefois, ce contre-pouvoir repose principalement sur la connaissance de l'informatique et de la conception du programme. D'autres groupes, des petits délinquants majoritairement, sont hors du système : du moment ou un délit même mineur est constaté, l'accès au Thunderhead est coupé. Ce pas de côté permet à certain.e.s de sortir volontairement de la surveillance, quand d'autres jouent sur la frontière et vivent un vrai deuil quand l'IA se tait pour elleux. Ce silence forcé est ce qui police la société : l'avantage d'une vie sans décision et sans pression vaut suffisamment pour maintenir l'immense majorité de la population dans la légalité.

Les faucheuses font ainsi leurs recherches et assurent tous les usages numériques de leur quotidien directement avec de la programmation – nous pouvons penser à des lignes de commandes – sans l'aide de l'intelligence artificielle qui accompagne pourtant tous les autres individus. Le hacking devient alors le meilleur moyen de contourner l'autorité en place : en modifiant le code soutenant l'autorité, cela la fera forcément varier. Car toute IA est dépendante de sa programmation. Les Unsavory, délinquants, s'organisent en communautés – qui sont exploré dans le troisième volume de la série, lorsque le thunderhead se tait et qu'eux seuls sont en mesure de vivre sans -- où la programmation prend une toute autre valeur : le numérique sert à augmenter les corps et les perception, à jouer avec ce qui est visible et accessible. La culture Modder est à son plein : il n'est plus question d'une solution universelle mais de patch et de hackathons, d'accès détourné aux réseaux de surveillance et de constitution de robots basiques. Certaines communautés se sont d'ailleurs fédérées autour d'un refus de cette instance gouvernante, s'employant à lui forcer des angles morts, à déconnecter ou remplacer des données etc.

Certaines zones, se constituent en zones franches, en ZAD, comme le Texas : des enclaves où les lois diffèrent, où le thunderhead n'a plus qu'une surveillance minime et très peu d'agentivité. Si elles sont d'abord présentées comme des lieux d'expérimentation sociale servant à tester les théories de l'IA, elles prennent peu à peu une coloration de Far West, d'enclaves permissives où la différence et la diversité peut s'exprimer sans être policée par une intelligence artificielle tentant de faire correspondre des personnes à des modèles (un peu comme ce qui peut être fait dans d'autres distopies pou Young Adult comme dans la série Divergent par exemple).

D'ailleurs, des anticipations moins futuristes comme *Une toile large comme le monde* d'Aude Seigne posent déjà cette question de façon très réaliste ; couper des câbles pouvant devenir la seule option pour sortir d'un réseau oppressant et hors de contrôle humain.

La question de modifier une conscience simplement en changeant le texte sur lequel elle repose est fascinante. Ici, les faucheuses doivent programmer des requêtes complexes pour mettre à jour des faits historiques, des enregistrements et des archives, leur permettant de faire la lumière sur les réalités de leur société, et ce, malgré les limitations

de la programmation de l'IA. En effet, ses concepteurs espéraient faire disparaitre les raisons de leur création, et mettre en avant le beau dans cette sur-intelligence synthétique plutôt que de conserver la mémoire des drames qui a réellement accompagné cette mise en place.

Heureusement, pourrions-nous dire, que les faucheuses ont bien eu accès à ces données, même si elles étaient partiellement cachées. En effet, la disponibilité est au cœur du propos : le Thunderhead est à même de manipuler l'immensité des données produites au quotidien par les sociétés humaines, mais aussi de les utiliser en second, en faire des statistiques pour dégager des tendances, par exemple, ou pour retrouver des archives ne semblant pas signifiantes pour une IA, mais l'étant pour un.e observateur.rice humain.e.

4. Une critique du numérique contemporain

Vous ne le savez peut-être pas, mais je suis candidate au doctorat en littérature numérique, et je m'intéresse aux gestes de lecture en environnement numérique (bien sur) : et donc plus largement à la littérature numérique et à notre rapport humain-machine dans une perspective culturelle. Ceci est particulièrement vrai dans mon travail quotidien à la chaire de recherche du Canada sur les écritures numériques, où je dois apprendre un peu de programmation et à manipuler divers outils numériques.
Comme souvent, la science-fiction peut nous servir à dégager une critique de notre réalité actuelle... et de la direction prise pour le futur immédiat.
En effet, le numérique suit le même chemin que ces intelligences artificielles : il est facile de laisser la simplicité gagner sans avoir à comprendre le fonctionnement réel du logiciel appliqué. N'oublions pas également que nous dépendons de plus en plus des environnements propriétaires : que ce soit un choix conscient ou non.

Ainsi, l'argument de la simplicité – de l'intuitivité – repose sur l'habitude à des codes communs et usuels, comme défini par Anthony Masure ou Alexander Galloway. Nous retrouvons dans scythe une opposition déjà usuelle au quotidien : entre usage en programmation -- en dur -- et manipulation d'interfaces graphiques et d'environnements ergonomiques.

Les interfaces guident les publications qui y sont faites ; comme toute technique, elles viennent avec des contraintes. Mais celles-ci étant numériques, elles nous sont moins évidentes, parce que moins pressantes. Or, ces contraintes calibrent les contenus publiés et donc en partie, du moins, les réflexions menées. C'est ce que théorise Marcello Vitali Rosati avec sa théorie de l'éditorialisation (2016, 2018, 2021).

Ces fictions nous permettent aussi d'entrevoir les possibles détournements qui peuvent venir bouleverser les règles de programmation mises en place.
L'exemple le plus frappant serait sans doute la place que prend la littérature dans les plateformes numériques : elle s'intègre peu à peu partout, sur des interfaces et des outils

non pensés pour la création. Sur les réseaux sociaux, le partage de texte à forte contrainte peut mener à des exercices d'écriture à contraintes particulièrement créatives. Le détournement tant des espaces que des syntaxes et cultures numériques permet d'ouvrir des espaces de créations infinis.

La culture numérique au sens large, dans une acception qui pourrait être synonyme de culture contemporaine, met à l'honneur la débrouillardise et le bricolage, remplaçant les clous des fab lab par des lignes de commandes et des hackathon associatifs. Penser le rapport à la culture numérique étant souvent penser le rapport à la culture libre et à l'ouverture des données, les intelligences artificielles et les décisions que nous prenons actuellement dans leur constitution sont un moteur décisionnel important. Leur apprentissage de corpus comme ceux de l'anthologia graeca pour les digital classics ne sont pas loin de l'apprentissage par inclusion des database des bibliothèques présentées dans Scythe… faut-il encore faire des choix éclairés et veiller à ne pas laisser de côté la réponse au fonctionnement des dites IA.

Bibliographie

Chambers, B. (2017) *A closed and common orbit*. London: Hodder.

Citton, Y. (2012) *Gestes d'humanités: anthropologie sauvage de nos expériences esthétiques*. Paris: Armand Colin (Collection: Le temps des idees).

Derien, M. (no date) *Intelligence artificielle : rêves et réalité – Anthropologie des interfaces Homme/Machine*. Available at: https://anthropo-ihm.hypotheses.org/2131 (Accessed: 11 April 2022).

Ewing, P. *et al.* (2019) *Neo Cab* [Nintendo Switch, iOS]. San Francisco: change agency. Available at: https://neocabgame.com.

Lacroix-De Sousa, S. (2022) 'L'approche par les risques adoptée par le projet de règlement sur l'Intelligence Artificielle de la Commission européenne', in *La régulation des algorithmes en matière bancaire et financière*. Orléans, France. Available at: https://hal.archives-ouvertes.fr/hal-03527560 (Accessed: 11 April 2022).

Masure, A. (2017) *Design et humanités numériques*. Paris: Éditions B42 (Collection Esthétique des données, 1).

Rumpala, Y. (2021) *Cyberpunk's not dead: laboratoire d'un futur entre technocapitalisme et posthumanité*. Saint-Mammès: le Bélial' (Parallaxe).

Seigne, A. (2017) *Une toile large comme le monde*. Carouge-Genève: Zoé.

Shusterman, N. (2016) *Scythe*. First edition. New York: Simon & Schuster BFYR (Arc of Scythe, 1).

Shusterman, N. (2018) *Thunderhead*. First Edition. New York: Simon & Schuster BFYR (Arc of a Scythe, book 2).

Shusterman, N. (2019) *The toll*. First edition. New York: Simon & Schuster Books for Young Readers (Arc of a Scythe, book 3).

Sillaume, G. (no date) *IA : la qualité des données ne suffit pas*. Available at: https://maisouvaleweb.fr/ia-la-qualite-des-donnees-ne-suffit-pas/ (Accessed: 11 April 2022).

Vint, S. (2021) *Science fiction*. Cambridge, Massachusetts: The MIT Press (The MIT Press essential knowledge series).

Vitali Rosati, M. (2016) 'Qu'est-ce que l'éditorialisation ?', *Sens Public* [Preprint]. Available at: http://sens-public.org/article1184.html (Accessed: 8 September 2019).

Vitali-Rosati, M. (2018) *On Editorialization: Structuring Space and Authority in the Digital Age*. Amsterdam: Institute of Network Cultures (Collection Theory on Demand).

Vitali-Rosati, M. (2020) 'Qu'est-ce que l'écriture numérique ?', *Corela. Cognition, représentation, langage* [Preprint], (HS-33). doi:10.4000/corela.11759.