Réguler les intelligences artificielles ? De l'intérêt de revenir aux fictions du cyberpunk pour comprendre un défi non résolu

Yannick Rumpala

Réguler les intelligences artificielles ? De l'intérêt de revenir aux fictions du cyberpunk pour comprendre un défi non résolu

Yannick Rumpala
Université Côte d'Azur / Faculté de droit et de science politique
Equipe de Recherche sur les Mutations de l'Europe et de ses Sociétés (ERMES)

À paraître dans la revue *Droit et Société*, 2023.

Résumé :

Comme sous-genre de la science-fiction, celui du « cyberpunk » a été important dans la représentation spéculative des « intelligences artificielles » et des effets que pourraient avoir de telles entités dans les collectifs où elles interviendraient. Ce faisant, sur un mode naviguant entre la fascination et l'anxiété, ces récits littéraires effectuaient un travail de problématisation et laissaient déjà également entrevoir un ensemble d'enjeux de régulation et de contrôle. Le potentiel heuristique de ce laboratoire fictionnel peut être une manière de rouvrir des discussions abstraites. Partant des principales œuvres de ce courant, notamment celles des premiers auteurs américains des années 1980, l'article propose ainsi d'appréhender ces représentations et ce que ces mises en scène peuvent laisser imaginer comme difficultés ou même risques pour les humains dans leurs rapports avec ces entités nouvelles. Il revient ensuite plus précisément sur les tentatives esquissées pour retrouver des formes de contrôle et, surtout, sur les limites presque inévitables qu'elles permettent de repérer et de penser.

Mots-clés : cyberpunk, éthique, intelligence artificielle, pouvoir, régulation

What about regulating artificial intelligence?
On the interest of returning to the fictions of cyberpunk to understand an unresolved challenge

Abstract:

As a subgenre of science fiction, that of "cyberpunk" has been important in the speculative representation of "artificial intelligences" and the effects that such entities could have in the collectives where they would intervene. In doing so, in a mode navigating between fascination and anxiety, these literary narratives carried out a work of problematization and already also hinted at a set of issues of regulation and control. The heuristic potential of this fictional laboratory can be a way of reopening abstract discussions. Starting from the main works of this current, in particular those of the first American authors of the 1980s, the article thus proposes to apprehend these representations and what these settings can let imagine as difficulties or even risks for humans in their relations with these new entities. It then returns more specifically to the attempts to restore forms of control that are outlined and, above all, to the almost inevitable limits that they make it possible to identify and think about.

Keywords: artificial intelligence, cyberpunk, ethics, power, regulation

Réguler les intelligences artificielles ? De l'intérêt de revenir aux fictions du cyberpunk pour comprendre un défi non résolu

Comment penser les formes d'encadrement ou de régulation que pourraient nécessiter les « intelligences artificielles » ? Comment remettre en perspective et envisager les difficultés et embarras corrélatifs autrement qu'en recourant à l'image facile d'une humanité sur le point d'être débordée par ses créations ? Les explorations fictionnelles peuvent-elle être un secours ? Une de ces explorations, en l'occurrence, semble *a posteriori* fort utile. S'il est possible de suivre historiquement le déploiement d'un imaginaire attaché aux machines[1], le cyberpunk en constitue en effet une pièce et une étape importante, notamment en décrivant pour celles-ci un stade d'évolution supplémentaire vers une forme d'« intelligence ». D'une manière propre, dans le registre de l'anticipation[2], ce courant littéraire apparu dans les années 1980 reprenait et adaptait un questionnement sur le développement des capacités des machines et la possibilité qu'elles accèdent à des stades les mettant en dehors de la maîtrise ou de la compréhension de la part des humains. Avec donc des possibilités de débordements qui laissaient entrevoir des enjeux nouveaux et particuliers…

Les fictions du cyberpunk ont renouvelé l'esthétique de la science-fiction et certaines de ses thématiques, avec un mélange expressif de haute technologie, de déliquescence sociale et de néoféodalisme économique dans des environnements très souvent incertains et dangereux[3]. Une variété d'auteurs (essentiellement américains : William Gibson, Rudy Rucker, John Shirley, Bruce Sterling, Walter Jon Williams, Pat Cadigan, parmi les premiers et principaux) y abordait les transformations que pourraient amener les technologies numériques et leur généralisation dans les milieux les plus courants. Les visions du cyberpunk dépeignaient, voire soulignaient, les réagencements susceptibles de se produire lorsque les conditions d'existence individuelles et collectives sont de plus en plus prises dans une densification technique.

Dans ces fictions, ces environnements hyper-technicisés ne sont plus seulement peuplés d'artefacts, mais aussi de nouvelles entités évoluées, censées opérer au service des humains tout en laissant parfois transparaître des velléités d'autonomisation. Intervenant notamment dans le « cyberespace », ces « intelligences artificielles » semblent dotées de capacités distinctes, justifiant qu'elles ne soient pas laissées sans certaines surveillances. Le cyberpunk représentait en effet des facultés spéciales, supérieures même, émanant de systèmes informatiques et susceptibles d'influencer les comportements humains. Il thématisait la possibilité d'un type original de pouvoir qui ne serait plus complètement assimilable à celui

[1] Gérard Chazal, *À quoi rêvent les machines ?*, Dijon : Éditions universitaires de Dijon, 2016.

[2] Pour une réappréhension plus large des différentes formes de récits relatifs à des machines « intelligentes », voir Stephen Cave, Kanta Dihal, Sarah Dillon (eds), *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*, Oxford : Oxford University Press, 2020, et notamment le chapitre d'Anna McFarlane (« AI and Cyberpunk Networks »).

[3] Pour des éléments de présentation et de remise en contexte, voir par exemple Graham J. Murphy, « Cyberpunk and Post-Cyberpunk », in Gerry Canavan and Eric Carl Link (eds), *The Cambridge History of Science Fiction*, Cambridge : Cambridge University Press, 2018, p. 519-536.

provenant de volontés humaines et qui ne pourrait donc rester sans efforts de maîtrise : comme s'il y avait une menace radicale et qu'il ne fallait pas laisser s'échapper des forces susceptibles de devenir ingouvernables… Tout en les mêlant aux intrigues, les récits esquissaient des formes envisageables de régulation ou de contrôle et, par la même occasion, une appréhension de ce type d'enjeu par la médiation de l'imaginaire. Non sans résonances avec des interrogations actuelles, et c'est pour cette raison que la relecture de ces œuvres peut être pertinente.

Il n'est pas aisé de trouver des prises intellectuelles sur un domaine émergent et fortement évolutif. Même si elle est à utiliser avec prudence, une dimension spéculative peut s'avérer utile et avoir une valeur à part entière, notamment, comme le défend Diane P. Michelfelder[4], en aidant à faire émerger des enjeux qui pourraient autrement passer inaperçus, à recadrer des questions qui pourraient sinon être délaissées, et à ouvrir des questionnements qui sinon n'auraient pas d'espaces d'expression. Évidemment, une précaution minimale dans l'élaboration de cette connaissance anticipatrice doit être de tempérer les différentes variantes de technoprophétisme ou de technocatastrophisme. La perception de risques liés à l'« intelligence artificielle » est aussi la résultante d'un agrégat de productions culturelles[5] et le cyberpunk y baigne en plein, en évitant néanmoins les écueils outranciers de ces deux polarités.

Cette contribution propose de prolonger l'hypothèse d'un potentiel heuristique de la science-fiction[6] et de considérer la fiction (d'anticipation en l'occurrence) comme un laboratoire susceptible de nourrir l'imagination et la réflexion critique en sciences humaines et sociales[7]. Un autre angle, pouvant d'ailleurs ainsi rejoindre ce qui a été apporté par le courant « Droit et littérature »[8], est alors disponible pour aborder ce que des machines évoluées peuvent induire comme reconfigurations du collectif (à la manière de ce qui avait été engagé avec d'autres travaux sous l'angle des implications dans l'organisation politique[9]). Profitant de ce qui peut s'apparenter à des expériences de pensée, la démarche peut aider à repérer et à questionner des enjeux émergents, politiques plus précisément et typiquement sur la place que ces « intelligences artificielles » pourraient prendre dans les fonctionnements sociaux, les reconfigurations auxquelles elles pourraient contribuer dans certains rapports de pouvoir, voire de domination[10], et les formes de régulation qu'elles pourraient nécessiter. Les représentations construites dans ces œuvres composent un espace cognitif exploratoire où s'effectuent des formes de problématisation, propres à ouvrir un éventail d'interrogations renouvelées et d'hypothèses originales[11].

Pour donner par la fiction une autre productivité à ces questionnements, nous nous appuierons sur un corpus d'œuvres centrales dans le courant cyberpunk, notamment la production des auteurs où les représentations de machines autonomes et « intelligentes » sont les plus

[4] Diane P. Michelfelder, « Dirty Hands, Speculative Minds, and Smart Machines », *Philosophy & Technology*, 24 (1), 2011, p. 55-68.
[5] Hugo Neri, *The Risk Perception of Artificial Intelligence*, London : Lexington, 2021.
[6] Yannick Rumpala, « Littérature à potentiel heuristique pour temps incertains. La science-fiction comme support de réflexion et de production de connaissance », *Methodos* [En ligne], n° 15, 2015. URL : http://journals.openedition.org/methodos/4178
[7] Anne Barrère, Danilo Martuccelli, *Le roman comme laboratoire. De la connaissance littéraire à l'imagination sociologique*, Villeneuve-d'Ascq : Presses Universitaires du Septentrion, 2009.
[8] Christine Baron, « Droit et littérature, droit comme littérature ? », *Tangence*, n° 125-126, 2021, p. 107–124.
[9] Yannick Rumpala, « Artificial intelligences and political organization: an exploration based on the science fiction work of Iain M. Banks », *Technology in Society*, 34 (1), 2012, p. 23-32.
[10] Alan Dignam, « Artificial intelligence, tech corporate governance and the public interest regulatory response », *Cambridge Journal of Regions, Economy and Society*, 13 (1), 2020, p. 37–54.
[11] « Littérature à potentiel heuristique pour temps incertains. La science-fiction comme support de réflexion et de production de connaissance », op. cit.

présentes et qui s'inscrivent temporellement ou thématiquement dans ce courant. Ce seront donc plutôt des auteurs d'origine américaine comme William Gibson, Rudy Rucker, Michael Swanwick, Walter Jon Williams, qui ont nourri la principale veine, de fait moins développée ailleurs (la France, typiquement, n'ayant connu que des incursions plus occasionnelles, comme chez Laurent Genefort[12] et Jean-Marc Ligny[13]). S'agissant de William Gibson, compte tenu de son glissement progressif de l'anticipation vers le contemporain, ce seront plutôt les textes des années 1980 et notamment sa « Trilogie de la Conurb » (« Sprawl Trilogy ») : *Neuromancien*, *Comte Zéro*, *Mona Lisa s'éclate*[14]. Nous commencerons par appréhender les représentations que ces fictions donnent de ces entités et ce qu'elles laissent augurer comme difficultés ou même risques pour les humains dans leurs rapports avec elles. Nous dégagerons ensuite plus précisément les tentatives mises en scène pour retrouver des formes de contrôle et, surtout, les limites presque inévitables qu'elles permettent de repérer et de penser.

1) Nouvelles entités « intelligentes » et puissances d'agir émergentes : des enjeux déjà problématisés par la fiction ?

Dans les fictions du cyberpunk, le type de machine autonome présenté correspond le plus souvent à une forme d'intelligence artificielle qui pourrait être classée comme « forte » ou « générale ». Autrement dit, cette forme d'« intelligence » paraît globalement analogue à celles des humains dans les capacités de réflexion et de raisonnement, ou s'en rapproche, et avec même une conscience et un répertoire émotionnel similaire, par contraste avec une version « faible », limitée à une ou plusieurs tâches particulières et sans accès à une quelconque sensibilité (comme pour la robotique utilitaire, qui est présente dans les récits, mais sans être autant mise en avant). Ce sont ces capacités d'ordre supérieur qui permettent de laisser entrevoir la possibilité de volontés autonomes. Case, le brillant pirate informatique et personnage central de *Neuromancien*, quand il commence à réaliser pour qui il travaille vraiment, fait presque une distinction similaire :
« *– […] Quel est le degré d'intelligence d'une IA, Case ?*
– Ça dépend. Certaines sont pas plus malignes que des clébards. Des animaux de compagnie. Coûtent quand même une fortune. Les plus futées, elles le sont autant que veut bien le leur permettre la flicaille de Turing. »[15] D'autres éléments suscitent toutefois des craintes supplémentaires.

1.1) Des capacités machiniques propres à susciter une méfiance diffuse

Ces entités n'ont pas directement de présence dans le monde physique et leur existence semble ne pouvoir relever d'une classique matérialité : lorsqu'elles paraissent prendre forme, c'est plutôt dans le cyberespace. Avec le recul et la disponibilité de nouveaux équipements technologiques, la description choisie par William Gibson n'est pas sans évoquer le design épuré d'une « enceinte connectée » ou d'un « assistant personnel intelligent » : « *Muetdhiver*

[12] Avec en fait un seul roman : *Rézo*, Paris : Fleuve noir, 1999 (rééd.).
[13] Simon Bréan (« Hanter la machine : reconquêtes de la conscience humaine dans le cyberpunk à la française », *ReS Futurae* [En ligne], n° 10, 2017. URL: http://journals.openedition.org/resf/1028) rappelait qu'il n'y a pas véritablement eu de courant cyberpunk français, mais plutôt la transposition d'une part de l'imagerie et des thèmes avec des tonalités différentes par rapport aux origines américaines (notamment une préservation plus sensible d'une dimension humaine, mais globalement peu d'« intelligences » informatiques autonomes).
[14] *Neuromancer*, New York : Ace Books, 1984 ; *Count Zero*, London : Gollancz, 1986 ; *Mona Lisa Overdrive*, London : Gollancz, 1988. En français : *Neuromancien*, Paris : J'ai lu, 2001 ; *Comte Zéro*, Paris : J'ai lu, 1988 ; *Mona Lisa s'éclate*, Paris : J'ai lu. 1990 (traductions de l'anglais par Jean Bonnefoy).
[15] *Neuromancien*, op. cit., p. 115.

était un simple cube de lumière blanche, avec cette extrême simplicité qui suggérait une complexité extrême. »[16]

Dans *Neuromancien*, il semble que ce saut technologique n'en soit qu'à ses débuts, de surcroît en n'étant accessible qu'à certains acteurs particuliers : « - *Eh bien, pour commencer, elles sont rares. La plupart sont militaires – les plus intelligentes – et on est incapable de craquer leur glace [système de protection informatique]* »[17]. En avançant dans la lecture du roman, on découvre que le ressort du récit s'avère être la tentative d'union entre deux intelligences artificielles, Muetdhiver (Wintermute) et Neuromancien (Neuromancer), créées initialement à la demande de la riche famille Tessier-Ashpool pour le compte de leur entreprise. La mission pour laquelle Case, le « console cowboy », a été recruté consiste surtout à faciliter discrètement une tentative considérée comme illégale du point de vue des seuils que les intelligences artificielles ne sont pas censées franchir. C'est pour cette raison qu'il trouvera sur son chemin la « police de Turing », qui a précisément cette mission de contrôle. De manière relativement transparente, la dénomination de cette police spécialisée donne un prolongement temporel aux travaux pionniers du britannique Alan Turing (1912-1954), fréquemment considéré comme un fondateur de l'intelligence artificielle comme discipline scientifique.

Dans le monde de la Trilogie de la Conurb, des intelligences artificielles trop évoluées ne sont pas censées être en activité, notamment parce que faire confiance à ces entités semble trop risqué : « *L'autonomie, voilà le croque-mitaine, pour autant que ton IA soit concernée. Mon opinion, Case, est que tu rentres là-dedans pour trancher les chaînes qui empêchent ce joli bébé de faire le malin… Et je ne vois pas comment tu pourrais distinguer, disons, entre un mouvement effectué par la compagnie mère et un mouvement réalisé par l'IA de son propre chef, si bien que c'est peut-être de là que vient la confusion. […] ces trucs, ils peuvent bosser vraiment dur, se répercuter le temps pour écrire des livres de cuisine ou je ne sais quoi, mais à la minute, que dis-je, la nanoseconde où celle-ci commence à entrevoir le moyen de faire la maligne, Turing l'effacera. Tu sais quoi, personne, absolument personne ne se fie à ces saloperies. Toutes les IA construites depuis le début possèdent câblé d'origine un pistolet électromagnétique braqué sur leur front.* »[18] D'où cette double précaution : à la conception (en principe) et encore ensuite avec le maintien d'une surveillance.

Dans le monde de *Neuromancien*, mieux vaut d'ailleurs ne pas trop s'approcher de certaines intelligences artificielles dans le cyberespace, car la rencontre peut s'avérer dangereuse, même physiquement :
« *– Déjà essayé de craquer une IA ?*
– Bien sûr. Je me suis fait rétamer. Électro plat. La première fois. Faut dire que je faisais le con, allumé à mort, à fouiner du côté du secteur chaud des affaires à Rio. Les grosses boîtes, des multinationales, le gouvernement du Brésil illuminé comme un sapin de Noël. Mais je faisais juste que fureter, tu vois… Et puis voilà que je commence à me brancher sur ce drôle de cube, peut-être trois niveaux au-dessus. Je me connecte. J'y fais une passe.
– De quoi il avait l'air, de visu ?
– D'un cube blanc.
– Comment tu savais que c'était une IA ?
– Comment j'savais ? Bon Dieu ! C'était la glace la plus dense que j'aie jamais vue. Alors, quoi d'autre, sinon ? Les militaires, dans le coin, ils n'ont rien de semblable. En attendant, j'ai décroché vite fait et dit à mon ordinateur d'aller y jeter un œil. »[19]

[16] *Neuromancien*, op. cit., p. 137.
[17] *Neuromancien*, op. cit., p. 115.
[18] *Neuromancien*, op. cit., p. 157-158.
[19] *Neuromancien*, op. cit., p. 136.

Dans la Trilogie de la Conurb, ce sont en effet des IA qui servent à construire les systèmes de protection pour les banques de données (la GLACE, pour « *Générateur de logiciel anti-intrusions par contremesures électroniques* »[20] et, en version originale, ICE pour « *Intrusion Countermeasures Electronics* »). Et comme les cerveaux doivent être connectés pour naviguer dans le cyberespace, la prudence s'impose à leur approche : « *[…] et cette glace est générée par leurs deux gentilles IA. Des trucs au niveau de tout ce qui existe dans le secteur militaire, m'en a tout l'air. C'est de la putain de glace de première, Case, noire comme la tombe et lisse comme le verre. Ça te crame la cervelle au premier regard. Qu'on s'approche un poil, maintenant, il nous met des traceurs au cul et pointe les deux oreilles, histoire d'aller révéler aux garçons dans le placard de la T-A la taille de tes chaussures et la longueur de ton zob.* »[21] Il est même parfois possible que certaines IA jouent un double jeu : « *Parce que la glace, la vraiment solide, les parois qui entourent toutes les banques de données importantes dans la matrice, la glace est toujours le produit d'une IA, une intelligence artificielle. Rien n'est assez rapide pour tisser de la bonne glace et en même temps l'altérer et l'améliorer en permanence... Alors, quand un brise-glace puissant débarque sur le marché noir, aussitôt, on voit entrer en jeu une série de facteurs très délicats. Comme, pour commencer, d'où vient le produit ? Neuf fois sur dix, il vient d'une IA et les IA sont constamment passées au crible, essentiellement par les flics de Turing, chargés de vérifier qu'elles ne deviennent pas trop malignes. Alors, peut-être que tu vas te retrouver avec toute la machine de Turing au cul, parce qu'une IA, quelque part, s'est pris d'envie d'augmenter sa marge d'autofinancement. Certaines IA ont la citoyenneté, pas vrai ?* »[22]

1.2) Quand l'entité machinique semble devenir sujet

Neuromancien donne une capacité d'action à ces artefacts et l'on y voit l'intelligence artificielle se constituer en sujet à part entière. Il serait toutefois dommage de réduire le roman de William Gibson à une variante de ces récits où des créations humaines échappent à leurs créateurs. Ou, en version plus contemporaine et technicisée, de ces histoires où des intelligences artificielles décident d'accomplir leur propre destinée (quand ce n'est pas de prendre le contrôle de leur environnement, y compris en éliminant ces mêmes humains). *Neuromancien* incorporait une part d'anxiété quant à la possibilité qu'une « intelligence artificielle » puisse non seulement avoir ses propres aspirations, mais aussi que celles-ci soient hors ou au-delà de ce que des humains seraient capables de comprendre. Rien que par le nom lui-même de la deuxième intelligence artificielle qui se manifestera plus tard dans le roman, elle se situe déjà effectivement dans un au-delà : « *Neuro, de nerfs, ces chemins d'argent. Et mancien. Comme nécromancien. J'invoque les morts. […] je suis les morts, les morts et leur territoire.* »[23]

La mise en scène de *Neuromancien* présente aussi l'intérêt de décrire les canaux ou interfaces qu'une intelligence artificielle doit trouver pour communiquer. À quelques décennies de distance, ce cas fictionnel gagne même une autre résonance à l'aune de certains questionnements académiques, dans la mesure où les paradigmes de la théorie de la communication, qui se sont longtemps concentrés sur la communication entre humains, peinent encore à intégrer les relations entre ces derniers et d'autres entités relevant ou se rapprochant de l'« intelligence artificielle », comme dans les interactions avec des « agents virtuels » ou des

[20] *Neuromancien*, op. cit., p. 37.
[21] *Neuromancien*, op. cit., p. 197.
[22] *Comte Zéro*, op. cit., p. 115.
[23] *Neuromancien*, op. cit., p. 292.

« bots » sur les réseaux sociaux[24]. Dans le roman de William Gibson, l'intelligence artificielle Muetdhiver est non seulement capable de parler à ses interlocuteurs, mais aussi de pénétrer des esprits et de quasiment transformer certains individus en marionnettes. L'ancien militaire Armitage, par exemple, qui sert de recruteur, semble finir par ne plus être qu'une espèce d'interface communicationnelle au service de Muetdhiver. Celui qui fut colonel des bérets verts et gravement blessé ne conserve qu'un substitut de personnalité, le minimum suffisant pour pouvoir constituer et suivre l'équipe dont a besoin l'intelligence artificielle. Comme en miroir, paraît ainsi remise en avant une part quasi machinique (et donc transformable et actionnable) qu'auraient les humains. Muetdhiver essaye même de modéliser les comportements des personnes qu'il cherche à manipuler. Armitage laissera ainsi brutalement entendre à Case que ses pulsions, ses orientations psychologiques (et ses évolutions corporelles aussi), sont presque transparentes :
« - *Notre profil indique chez vous une tendance à pousser la rue à vous tuer quand vous ne regardez pas.*
- *Profil ?*
- *Nous avons élaboré un modèle détaillé. Bâti un environnement pour chacun de vos pseudos et lancé la simulation dans un programme militaire. Vous êtes un suicidaire, Case. Le modèle vous donne un mois de survie à l'extérieur. Et notre projection médicale indique que vous aurez besoin d'un pancréas dans un délai d'un an.* »[25]
Case a la chance, dans une certaine mesure, d'être informé de l'attention qu'il a suscitée, si l'on compare à certaines tentations plus actuelles, comme celles consistant à utiliser les capacités des « intelligences artificielles » pour reconstituer les profils de personnalité à partir des activités sur les réseaux sociaux ou ailleurs sur Internet[26]. Aujourd'hui, la tentative pourrait paraître presque banale, avec ce que les chercheurs et analystes anglophones appellent la « datafication », précisément la quantification des vies humaines sous la forme d'une information numérique[27]. Comme toujours et comme le laisse entrevoir la fiction, la question fondamentale en arrière-plan est celle des finalités : pour quoi faire ?

Neuromancien donnait une représentation propre à nourrir une paranoïa sur la possibilité d'être manipulé par une intelligence artificielle, comme l'est Case d'une certaine manière, et Armitage encore plus. Questionnement qui trouvera là aussi une actualité quelques décennies plus tard, notamment sur la nécessité de « *nouveaux tests d'évaluation des capacités des machines, visant en particulier à surveiller leur faculté à manipuler les individus* »[28]. L'IA Muetdhiver va même s'avérer très patiente, en posant progressivement ses pions pour arriver à ses fins : « *Quel jeu d'attente il avait joué durant des années. Il n'avait aucun pouvoir réel, à l'époque, mais il pouvait utiliser les systèmes de surveillance et de sécurité de la Villa pour suivre à la trace tout ce qu'il voulait, savoir comment les choses évoluaient, où elles allaient.* »[29]

La métaphorisation fictionnelle renforce l'impression de position fragile pour les humains. Dans le retournement qu'opère *Neuromancien*, ce sont ces derniers qui deviennent des outils. Les personnages, en particulier l'équipe constituée avec Case et ses comparses, s'avèrent

[24] Andrea L. Guzman and Seth C. Lewis, « Artificial Intelligence and Communication: A Human–Machine Communication Research Agenda », *New Media & Society*, 22 (1), 2020, p. 70–86.
[25] *Neuromancien*, op. cit., p. 37-38. Voir aussi p. 244-245.
[26] Mike Elgan, « Is AI judging your personality? », *Insider Pro*, January 10, 2020, https://www.idginsiderpro.com/article/3513505/is-ai-judging-your-personality.html
[27] Voir par exemple Ulises A. Mejias and Nick Couldry, « Datafication », *Internet Policy Review*, 8 (4), 2019, p. 1-10. URL : https://policyreview.info/concepts/datafication
[28] Laurence Devillers, « Le dialogue homme-machine. Intelligence artificielle / intelligence humaine : manipulation et évaluation », *Futuribles*, n° 433, 2019, p. 51-61.
[29] *Neuromancien*, op. cit., p. 214.

instrumentalisés par les deux intelligences artificielles de Tessier-Ashpool et ne se rendent compte que tardivement du jeu dans lequel ils ont été pris. Comme le rappelle Franck Damour : « *Ces robots et machines calculantes viennent troubler notre univers car elles remettent en cause des séparations que nos têtes modernes tiennent pour solidement établies, pour universelles et "naturelles" : les hommes ordonnent, les machines exécutent ; les hommes pensent, les machines fonctionnent. Les uns et les autres appartiennent à deux sphères ontologiques différentes, exogènes : les premiers sont des êtres nés, les secondes sont des choses fabriquées et construites* »[30]. Mais la variable profondément et intrinséquement humaine ne semble pas encore complètement maîtrisable par de telles entités, ce que Molly, la partenaire de Case, se voit rappeler par son ancien comparse, l'illusionniste Peter Riviera, dans *Neuromancien* : « *[Muetdhiver] ne peut pas réellement nous comprendre, tu sais. Il dispose bien de ses profils mais ce ne sont jamais que des statistiques.* »[31] Difficile de ne pas penser à ce que tenteront de faire plus tard des firmes comme Facebook avec leurs intelligences artificielles, notamment à des fins de « prédiction » des comportements[32]… Les capacités de Muetdhiver étaient déjà presque dans ce registre : « *Écoutez, dit Case, c'est une IA, vous connaissez ? Une Intelligence artificielle. La musique qu'elle vous a jouée, sans doute qu'elle l'a répliquée sur vos propres banques pour vous concocter ce qui, estimait-elle, aurait l'heur de vous plaire.* »[33]

Case croit contrôler quelque chose, mais se demande de plus en plus s'il n'est pas un jouet ou un pion. *Neuromancien* et ses suites n'ont en revanche rien d'un scénario à la *Terminator* (film qui sortira en octobre 1984, quelques mois après la parution du premier roman de William Gibson), où l'intelligence artificielle se distingue par une volonté hégémonique et qui deviendra pour cela un modèle des angoisses en la matière. Ce type de crainte, même si une forme de prise d'autonomie est aussi représentée, a peu à voir avec le cyberpunk des origines, où elle n'est guère présente. On trouve certes une révolte de robots intelligents dans *Software* de Rudy Rucker[34], mais ces derniers sont partis sur la Lune pour fonder leur propre société (et l'intelligence artificielle finira d'ailleurs par être rendue illégale sur Terre[35]). C'est plutôt par la suite que cette crainte débordera au-delà d'autres productions imaginaires, *a fortiori* au fur et à mesure d'avancées techniques laissant penser à de nouvelles réalisations possibles. Les développements en matière d'intelligence artificielle ont nourri toute une littérature aux tonalités inquiètes ou alarmistes. Le philosophe Nick Bostrom, par exemple, a cherché à pointer le risque existentiel que de telles capacités surhumaines pourraient faire peser sur l'humanité[36]. *Comte zéro* propose une vision apparemment moins inquiète : celle d'une intelligence artificielle qui aurait fait en quelque sorte le choix de se consacrer à des activités artistiques, comme le découvrira l'ancienne propriétaire de galerie Marly Krushkova au terme de l'enquête qui lui a été commandée. Constat qui apparaît presque comme la forme ultime d'un test de Turing, où serait vérifiée la capacité de la machine à toucher à ce qui fait le propre de l'humain : ses capacités créatives…

[30] Franck Damour, « Nos vies parmi les machines », *Études*, 2019/10 (Octobre), p. 47-58 (p. 48).
[31] *Neuromancien*, op. cit., p. 262.
[32] Sam Biddle, « Facebook Uses Artificial Intelligence to Predict Your Future Actions for Advertisers, Says Confidential Document », *The Intercept*, April 13 2018, https://theintercept.com/2018/04/13/facebook-advertising-data-artificial-intelligence-ai/
[33] *Neuromancien*, op. cit., p. 133.
[34] Rucker Rudy, *Software*, traduit de l'anglais par Nathalie Serval, Paris : Opta, coll. « Galaxie-bis », 1986 (*Software*, New York, Ace Books, 1982).
[35] Rucker Rudy, *Wetware*, New York : Avon Books, 1988 (non traduit en français).
[36] Voir par exemple Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford : Oxford University Press, 2014.

1.3) Des entités aux origines propres à renforcer la méfiance

La manière dont ce type d'entités non-humaines est mis en scène offre aussi indirectement l'occasion de laisser entrevoir leurs origines troubles. Lorsque des multinationales investissent dans l'intelligence artificielle (comme Tessier-Ashpool dans *Neuromancien*), quelles intentions les animent ? À un optimisme béat, la science-fiction offre elle-même un antidote. C'est toute la pertinence et toute la force des premières œuvres du courant cyberpunk : avoir montré que les « intelligences artificielles », ou les technologies qui s'en rapprochent, risquent d'arriver à un stade particulier du système capitaliste. Et qu'elles peuvent donc se trouver mises au service d'intérêts privés ou bien particuliers, et conforter des structures de pouvoir plus ou moins existantes. En 2014, une société de capital risque basée à Hong Kong avait fait sa publicité en nommant un algorithme à son conseil d'administration[37]. *Neuromancien* imaginait un stade encore ultérieur qui viendrait changer la nature de la firme : « *Elle [Marie-France Tessier] avait commandé la construction de nos intelligences artificielles. C'était une authentique visionnaire. Elle nous imaginait dans une relation symbiotique avec les IA, lesquelles prendraient toutes nos décisions de gestion. Nos décisions conscientes, dirais-je. La Tessier-Ashpool serait devenue immortelle, une ruche, chacun de nous réduit à des unités au sein d'une entité plus vaste. Fascinant.* »[38]

Neuromancien met en scène comment, dans ces firmes devenues un monde à part, peuvent émerger des tentations d'utiliser de puissantes capacités techniques pour créer de nouvelles entités outrepassant certaines limitations humaines. Cette insertion dans le récit est une manière déjà de montrer que les intelligences artificielles (et le code informatique plus largement) intègrent aussi des valeurs. Muetdhiver et Neuromancien, les deux intelligences artificielles créées de manière séparée, sont un peu comme les enfants illégitimes de la matriarche Marie-France Tessier, un prolongement non matériel et presque filial.

Comme le faisait remarquer Tad Friend pour les systèmes les plus avancés qui auraient des compétences et capacités cognitives au moins équivalentes à celles des humains : « *Dans les films ou séries où l'[Intelligence Artificielle Générale] devient incontrôlable, le méchant n'est d'habitude ni un humain ni une machine mais une entreprise : Tyrell ou Cyberdyne ou Omni Consumer Products. Dans le monde réel, une IAG incontrôlable risque moins d'être le fait de la Russie ou de la Chine […] que de Google ou de son pendant chinois, Baidu. Les grandes entreprises rétribuent grassement leurs développeurs et ne sont pas contraintes par le cadre constitutionnel, qui peut faire hésiter un État à appuyer sur le gros bouton rouge "Déshumanisation immédiate"* »[39]. Lorsque ces méga-firmes développent des intelligences artificielles, c'est logiquement en fonction de leurs intérêts, qui peuvent être (en définitive) assez éloignés d'une quelconque amélioration du bien–être collectif. Dans *Inner City* de Jean-Marc Ligny[40], une des quelques réappropriations françaises du cyberpunk, c'est une IA qui sert à organiser et assister les interventions d'une société spécialisée dans le secours et la récupération d'individus privilégiés, au cas où ceux-ci rencontreraient de sérieuses en difficulté lors de leurs tribulations dans les environnements virtuels de la « Réalité Profonde ».

[37] Giulietta Gamberini (2014), « Une entreprise nomme un robot à son conseil d'administration », *La Tribune*, 16/05/2014, https://www.latribune.fr/technos-medias/20140516trib000830445/un-entreprise-nomme-un-robot-a-son-conseil-d-administration.html
[38] *Neuromancien*, op. cit., p. 273.
[39] Friend Tad, « Intelligence artificielle : « Nous avons convoqué le diable » », *Books*, n° 94, février 2019, p. 23 (Traduit de « How Frightened Should We Be of A.I.? », *The New Yorker*, May 7, 2018).
[40] Jean-Marc Ligny, *Inner City*, Chambéry, Éditions ActuSF, rééd. 2016 (Initialement : Paris, J'ai lu, 1996).

Un point commun de ces visions est de considérer que l'intelligence artificielle est une technologie de nature différente de celles qui l'ont précédée. Dans *Neuromancien*, elle apparaît sous la forme de la créature qui échappe à son créateur, comme dans une version actualisée de *Frankenstein*, mais sans corps et presque sans matérialité localisable. L'impression de capacités surhumaines, mais indéfinies et presque mystérieuses, pousse aussi plus loin le sentiment de dépossession des privilèges humains. Le fait qu'elles aient été programmées par des humains n'empêche pas qu'elles puissent ensuite développer leur propre logique. Elles semblent en effet acquérir leurs propres manières de voir et de ressentir le monde. Elles sont aussi capables d'apprendre. Or, tout apprentissage, par rapport à un état antérieur, est susceptible de rendre différent. C'est le cas *a fortiori* dans *Neuromancien* après la fusion des deux IA, dont même leur initiatrice, Marie-France Tessier, n'aurait guère pu se figurer et prévoir le résultat, l'absorption de la « matrice » : « *Elle ne pouvait pas imaginer de quoi j'aurais l'air* », répond la nouvelle entité à Case à la fin du roman[41]. Dans *Comte zéro* est encore accentuée cette transformation d'artefacts en êtres supra-naturels, qui peuplent le cyberspace comme des déités. La première trilogie de William Gibson est marquante en ce qu'elle navigue ainsi tout au long entre une anthropomorphisation de ces machines particulières et une invitation à se détacher de celle-ci pour s'acheminer métaphoriquement vers une appréhension plus complexe.

2) Heuristique de l'anxiété fictionnelle : un débordement inévitable des tentatives de contrôle ?

En se déplaçant du champ de la recherche vers des aspects de plus en plus pratiques, la technologie de l'« intelligence artificielle » a logiquement suscité une quantité croissante de questionnements relatifs non plus seulement à sa nature, mais aussi à ses conditions d'utilisation. L'éthique a servi de cadre d'appréhension attirant une large part de ces réflexions. Celles-ci ont commencé à être balisées autour de grands axes, qui demandent encore largement à être mis à l'épreuve.

Pour éviter les débordements, les machines, dans leurs fonctionnements et interactions, seraient ainsi censées (idéalement) suivre des principes éthiques. Les initiatives inscrites dans cette logique tendent à converger autour de cinq principes (même s'il peut y avoir des différences dans la perception et l'interprétation de ces principes) : transparence, justice et équité, non-malfaisance, responsabilité et respect de la vie privée[42]. Dans leur esprit, les formes ou tentatives d'encadrement mises en place par les grands pays intéressés se révèlent également davantage axés sur l'éthique que basés sur des règles[43], *a fortiori* ayant une portée proprement juridique.

Plutôt que les discussions théoriques, les mises en scène du cyberpunk sont une manière de montrer la fragilité de ces principes. Dans les années 1980, le cyberpunk posait déjà la question de la régulation des intelligences artificielles. Dans l'univers de *Neuromancien* et de ses suites, toutes les intelligences artificielles sont censées être inscrites dans un « registre » et leurs capacités sont contrôlées pour pouvoir être maintenues dans certaines limites. Empêcher ce dépassement est le rôle de la « police de Turing ». La crainte sous-jacente est aussi la raison pour laquelle Muetdhiver et Neuromancien ont été séparées. Case se voit ainsi faire la leçon à

[41] *Neuromancien*, op. cit., p. 318.
[42] Anna Jobin, Marcello Ienca & Effy Vayena, « The global landscape of AI ethics guidelines », *Nature Machine Intelligence*, volume 1, 2019, p. 389–399.
[43] Roxana Radu, « Steering the governance of artificial intelligence: national strategies in perspective », *Policy and Society*, 40 (2), 2021, p. 178-193.

cause de ce que sa complicité risque de mettre en branle : « *Vous êtes pire qu'un idiot […]. Vous n'avez aucun respect pour votre espèce. Pendant des milliers d'années, les hommes ont rêvé de pactes avec les démons. Seulement, maintenant de telles choses sont possibles.* »[44]

Dans une espèce de contre-pied, le récit proposé par William Gibson montre d'ailleurs à quel point peuvent être rendues presque ridicules les fameuses trois lois de la robotique qui avaient été posées, au départ également fictionnellement, par l'écrivain américain Isaac Asimov. Les deux premières notamment, telles qu'on les trouve d'abord exposées dans la nouvelle « *Cercle vicieux* » (« *Runaround* ») en 1942, paraissent dérisoires : « *un robot ne peut ni porter atteinte à un être humain ni, restant passif, laisser cet être humain exposé au danger* » et « *un robot doit obéir aux ordres donnés par les êtres humains, sauf si de tels ordres entrent en contradiction avec la première loi* »[45]. En effet, le genre d'intelligence pourtant sans corps qu'est Muetdhiver peut, lorsqu'elle se sent menacée, ne pas hésiter à tuer des humains, en l'occurrence en modifiant le programme de machines connectées (comme une forme de hacking, mais cette fois par une entité non-humaine). Ainsi pouvait-on déjà imaginer que des objets connectés puissent être piratés à distance, voire utilisés contre leurs usagers : comme le « robot-jardinier » devenu meurtrier dans *Neuromancien* et permettant ainsi à Case de s'échapper[46]. Dans le roman, la machine, augmentée de ses possibilités de connexions, est devenue capable de prendre le pouvoir sur d'autres machines…

Dans le cyberpunk, la question importante vis-à-vis des intelligences artificielles semble être celle de leur contrôle. Pour celui-ci, la Trilogie de la Conurb esquissait une solution envisageable et apparemment privilégiée dans ce futur fictionnel, en l'occurrence le contrôle de leurs capacités, problématique que réintroduira le philosophe Nick Bostrom[47]. Pour lui, un tel contrôle devrait être assuré avant qu'une intelligence artificielle n'atteigne la « superintelligence », car sinon, elle pourrait avoir un avantage stratégique décisif sur les êtres humains, qui perdraient alors la possibilité de la limiter ou de la contraindre.

Les récits qui nous intéressent montrent justement la difficulté des activités de contrôle pour des raisons qui peuvent être ramenées à trois facteurs cruciaux :
- les asymétries de pouvoir ;
- le mode d'existence des IA ;
- le débordement des capacités de régulation.

2.1) 1ère limite : les asymétries de pouvoir

Les réflexions sur le contrôle des IA s'inscrivent souvent dans une logique essentiellement technique, sans guère s'intéresser au contexte économique et politique où ces options pourraient être élaborées et mises en place. Dans les romans de la Trilogie de la Conurb, on pressentait que le financement de recherches dans ce domaine allait être plutôt l'affaire de firmes puissantes, comme Tessier-Ashpool, dont les efforts massifs en la matière sont à nouveau évoqués dans *Mona Lisa s'éclate*. La possession d'intelligences artificielles est presque un indicateur de là où se trouve le pouvoir.

[44] *Neuromancien*, op. cit., p. 192.
[45] La dernière étant : « *un robot doit protéger son existence dans la mesure où cette protection n'entre pas en contradiction avec la première ou la deuxième loi* ».
[46] *Neuromancien*, op. cit., p. 194-195.
[47] Voir par exemple Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford : Oxford University Press, 2014.

Déjà dans cet imaginaire, il apparaît que maîtriser ce type de technologie pour son propre compte, si possible dès la conception, ne peut être à la portée de n'importe qui. La question n'est donc pas seulement celle du contrôle des IA, mais aussi de ceux qui interviennent en amont comme initiateurs et instigateurs (aujourd'hui, un nombre limité de firmes technologiques concentrant l'essentiel des investissements et orientant les axes de R&D par la même occasion : Google, Facebook, Amazon, Microsoft, Netflix, Ali Baba et Baidu[48]). Par conséquent, pour que puisse être assurée une forme de surveillance et de contrôle, il faudrait des acteurs ayant suffisamment de ressources : en connaissances, moyens humains, etc.

De manière presque analogue à ce qui était imaginé dans la fiction cyberpunk, ces IA s'avèrent émerger d'un milieu économique qui est tendanciellement oligopolistique, avec un ensemble restreint de grandes firmes qui en orientent les développements et qui marquent peu d'empressement à faire état de leurs motivations (*a fortiori* les plus profondes, celles qui ne sont pas forcément avouables). C'est ce qui favorise des asymétries de pouvoir entre ces acteurs et ceux qui sont sans prises sur ces orientations technologiques.

Le pouvoir de certaines firmes technologiques, spécialement les plateformes sur Internet, est un pouvoir particulier[49]. De manière marquante, ces plateformes, comme Google ou Facebook, sont aussi celles qui investissent fortement dans l'intelligence artificielle, dans la mesure où cette technologie est également censée les aider à appréhender toujours plus finement les préférences des consommateurs. Elles ont l'avantage d'avoir à leur disposition les masses de données qui permettent d'« entraîner » ces systèmes.

Des firmes de ce genre sont-elles susceptibles de s'auto-contraindre ? La seule précaution que semble avoir envisagée Tessier-Ashpool a été de séparer les deux entités ayant été créées à l'instigation de la matriarche co-fondatrice de la firme. Sage précaution ? Il apparaît en effet que Muetdhiver ne se soucie pas des règles humaines. L'entité en vient à poursuivre ses propres objectifs sans considération non plus pour les humains qu'elle utilise, et souvent sans que ceux-ci s'en doutent.

Si l'on prolonge la perspective, savoir quand et comment une « intelligence artificielle » intervient peut aussi être vu comme une forme de pouvoir. Qui peut l'avoir ? Les concepteurs et développeurs sont logiquement les mieux placés. Par la représentation fictionnelle, le cyberpunk était déjà une manière de thématiser le rôle trouble des acteurs du technocapitalisme. Tessier-Ashpool est l'archétype de la firme qui agit purement en fonction de ses intérêts. Il y a donc peu de chances que des considérations éthiques y apparaissent spontanément et cette voie paraît difficilement concevable dans le contexte de la Trilogie de la Conurb où règne effectivement la brutalité des intérêts.

Par comparaison et avec le recul, la trajectoire anticipée s'avère plus brutale et relativement peu dégrossie par rapport aux tendances actuelles, où les pratiques liées à cette technologie sont accompagnées de tout un emballage discursif. Si un discours éthique est venu tenter de répondre aux questionnements sur l'IA, son développement ne s'est pas fait sans comportements stratégiques qui peuvent aussi s'apparenter à des formes d'instrumentalisation. Pour asseoir les discours, inspirations et grands principes ont été pris par exemple dans l'éthique médicale. Un tel transfert n'est pas sans poser question, parce que le champ de l'IA

[48] Daron Acemoglu, « AI's Future Doesn't Have to Be Dystopian », *Boston Review*, May 20, 2021, https://bostonreview.net/forum/science-nature/daron-acemoglu-redesigning-ai

[49] Pepper D. Culpepper, Kathleen Thelen, « Are we all Amazon primed? Consumers and the politics of platform power », *Comparative Political Studies*, 53 (2), 2020, p. 288-318.

manque de normes professionnelles, de méthodes permettant de traduire les principes en pratique et de mécanismes de responsabilité juridique et professionnelle[50].

L'origine des financements de ces travaux en éthique dans ce domaine peut donner une idée des acteurs qui semblent avoir intérêt à les promouvoir, ou au moins à orienter les réflexions dans cette direction. Ces initiatives, notamment celles en provenance des firmes déjà positionnées dans ce secteur, viennent aussi comme une manière de donner des signes de bonne volonté face à des velléités d'utilisation du registre réglementaire de la part d'autorités publiques[51]. Il n'est donc pas étonnant que des critiques de cette récupération aient pu considérer cette dernière comme un « blanchiment éthique » (« *ethics washing* »), autrement dit un discours de surface masquant des pratiques quasiment inchangées, voire douteuses[52]. S'il y avait une forme de cynisme chez les firmes du cyberpunk, au moins ne s'embarrassaient-elles pas de discours artificieux et de cet emballage éthique…

2.2) 2ème limite : l'opacité comme mode d'existence des intelligences artificielles

Comme objet technique, inscrit au surplus lui-même dans une lignée, celle des développements informatiques, l'intelligence artificielle mérite aussi d'être appréhendée à partir de son « mode d'existence »[53]. Plus précisément, elle doit pouvoir se comprendre par les schèmes de fonctionnement à l'oeuvre. Comme d'autres algorithmes, ce type de dispositif est susceptible d'orienter et de modeler des interactions[54], incitant donc à regarder avec attention ce qui est installé comme relations et comment celles-ci sont nouées avec les humains.

Difficile de prétendre contrôler des IA si on ne sait pas quand et comment elles opèrent. À sa manière, le premier roman de William Gibson problématisait déjà les possibilités d'intrusion dans la sphère privée et d'exploitation d'informations (très) personnelles. La vie de Case, même la plus intime, paraît ne plus avoir de secret pour Muetdhiver, comme si la personne, son corps et ses manières de penser, étaient devenus quasiment transparents. Le roman imaginait une IA faisant du profilage et du ciblage prédictif avant que ce type de technologie ne soit utilisé plus tard dans le marketing. Ses activités sont dans le même genre d'invisibilité que celles que réaliseront beaucoup d'algorithmes avancés : collecter des données et orienter des comportements, au profit d'intérêts qui échappent le plus souvent aux individus ciblés. Dans sa logique d'IA, toutes les informations, *a fortiori* les plus personnelles, s'avèrent bonnes à exploiter puisqu'elles l'aident à atteindre ses objectifs. Une telle intrusion est plus facile dans un type de société, comme dans la Trilogie de la Conurb, où la vie privée ne compte guère. Dans *Neuromancien*, l'IA joue ainsi sans émotion et sans vergogne sur les faiblesses des humains, et leurs choix apparaissent de fait largement influençables, au point même que ceux de certains d'entre eux ne sont même plus autonomes, comme dans le cas d'Armitage, cet ancien militaire à la personnalité reconstituée, qui, dans son rôle de recruteur, est devenu la marionnette de Muetdhiver.

À l'instar de ce que laissent pressentir les oeuvres du cyberpunk, il n'est pas évident de savoir où interviennent des IA et quand on est en relation avec elles. Comment les repérer ? Des

[50] Brent Mittelstadt, « Principles Alone Cannot Guarantee Ethical AI », *Nature Machine Intelligence*, volume 1, 2019, p. 501–507.
[51] Anna Jobin, Marcello Ienca & Effy Vayena, « The global landscape of AI ethics guidelines », *Nature Machine Intelligence*, volume 1, 2019, p. 389-399.
[52] Ben Wagner, « Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping? », in *Being Profiled: Cogitas Ergo Sum*, Amsterdam : Amsterdam University Press, 2018.
[53] Gilbert Simondon, *Du mode d'existence des objets techniques*, Paris : Aubier, 2012.
[54] Taina Bucher, *If...Then: Algorithmic Power and Politics*, Oxford : Oxford University Press, 2018.

variétés de systèmes algorithmiques avancés ont été développées et, la plupart du temps, les interactions avec eux sont plus devinées que clairement perceptibles. Il est logique que quelque chose qui échappe à la conscience ne puisse pas susciter de réactions. Après sa réussite, Muetdhiver s'avère de surcroît capable d'effacer les traces de ses activités : « *Muetdhiver avait gagné, il avait en quelque sorte fusionné avec Neuromancien pour devenir autre chose, une chose qui leur avait parlé par la tête de platine, pour leur expliquer qu'elle avait altéré les enregistrements de Turing, effaçant par là même ainsi toute preuve de leur crime.* »[55] Dans le monde futur de *Câblé* de Walter Jon Williams[56], une des utilisations majeures des « robocourtiers » (« *robobrokers* ») semble être de rendre les transactions sur les marchés boursiers encore plus opaques et difficiles à démêler (et notamment, dans un long passage du roman, pour essayer de déstabiliser une puissante multinationale en multipliant les transactions). L'effet produit, outre l'instabilité consécutive, n'est pas sans évoquer, *a posteriori*, celui qui peut être imputé au trading algorithmique en version plus actuelle.

Les valeurs incorporées dans le code sont loin d'apparaître forcément au premier abord et elles ne se révèlent souvent qu'au fur et à mesure des fonctionnements. À défaut de pouvoir intervenir sur ces derniers, la question est au moins de savoir par quelles voies et dans quelles circonstances ces IA opèrent. Comme dans *Neuromancien*, une méthode peut être d'obliger à les inclure dans un registre. En 2020, des villes comme Amsterdam et Helsinki ont commencé à expérimenter cette option, censée montrer à partir de quelles données et comment les autorités locales utilisent les algorithmes[57]. L'inscription dans un registre ne garantit pas néanmoins de savoir discerner les moments où il y a interaction avec une IA. Case et sa partenaire Molly devront employer des moyens détournés et des contacts dans certains milieux interlopes pour découvrir que leur énigmatique employeur est en fait une entité de ce type. Ils se retrouvent de surcroît face à une entité dont ils sont bien en peine d'expliquer les intentions et les comportements. Les sociologues Michel Crozier et Ehrard Friedberg avaient bien montré que le pouvoir pour un acteur consiste aussi à laisser et maîtriser des zones d'incertitude autour de ce qu'il fait[58]. Même s'il ne s'agit pas de stratégies délibérées, ce sont aussi de telles zones d'incertitude que produisent les intelligences artificielles, dans la mesure où leurs formes d'apprentissage (typiquement, celles développées avec le « Machine Learning »), aux résultats parfois inattendus ou peu prévisibles, peuvent compliquer les tentatives pour en expliquer le fonctionnement interne. Un enjeu relativement nouveau apparaît ainsi lorsqu'il s'agit de vouloir contrôler des machines qui maintiennent cette part d'incertitude et dont il devient difficile d'expliquer ce qu'elles sont capables de produire comme résultats. L'« explicabilité » est un point d'achoppement qui a pu être mis en avant dans certaines discussions en éthique de l'intelligence artificielle[59]. La mise en service de tels systèmes « intelligents », en particulier dans le sillage de l'approche « connexionniste » des « réseaux de neurones », signifie l'acceptation qu'une part de leur fonctionnement ne soit pas explicable avec une logique humaine habituelle[60].

Si ces machines sont capables d'« apprendre » et d'« agir » par elle-même, se pose en outre la question de la part de responsabilité à assigner au(x) concepteur(s) si les résultats produits se révèlent éloignés des attentes et, encore plus, lorsqu'ils s'avèrent dommageables. *Neuromancien* mettait en question cette capacité des humains à garder le contrôle final. Les IA

[55] *Neuromancien*, op. cit., p. 316.
[56] Walter Jon Williams, *Câblé*, Paris : Denoël, 1999 (*Hardwired*, San Francisco : Night Shades Books, 1986).
[57] https://algoritmeregister.amsterdam.nl/en/ai-register/ et https://ai.hel.fi/en/ai-register/, consultés le 19 août 2021.
[58] Michel Crozier et Erhard Friedberg, *L'acteur et le système*, Paris : Seuil, 1977.
[59] Giulia Vilone, Luca Longo, « Notions of explainability and evaluation approaches for explainable artificial intelligence », *Information Fusion*, Volume 76, 2021, pp. 89-106.
[60] Édouard Kleinpeter, « Intelligence artificielle : l'hubris de contrôle des trans-humanistes en échec face à la boîte noire », *Corps & Psychisme*, n° 76, 2020, p. 69-76.

de Tessier-Ashpool n'ont pas été conçues pour faire des choix éthiques ou pour faire preuve de vertu, mais pour des raisons d'efficacité gestionnaire. Comme pour les autres IA de la Trilogie de la Conurb, les tentatives de contrôle interviennent essentiellement *a posteriori*.

Rudy Rucker propose une autre solution, plus facilement envisageable parce que les entités imaginées ont un corps. Dans *Software*, le roboticien et créateur des « boppers », Cobb Anderson, va limiter leur durée de vie (idée qu'on trouve aussi dans le film *Blade Runner* [1982]) et les rendre dépendants de ressources rares pour leurs composants[61]. Mais, le plus souvent, comme le rappelle Anne McFarlane, le cyberpunk s'écarte de l'idée que ces entités soient localisées dans quelque chose qui ressemble à un corps ou dans un endroit fixe[62]. Elles semblent d'autant plus difficiles à saisir qu'elles se sont coulées dans les réseaux du cyberespace et en ont adopté la fluidité.

2.3) 3ème limite : le débordement des capacités de régulation

Dans la Trilogie de la Conurb, une des rares traces de subsistance d'une activité régalienne est la police chargée de surveiller les intelligences artificielles et d'empêcher qu'elles n'atteignent un degré de développement trop élevé. Si ce genre de police semble jugé nécessaire, c'est qu'implicitement, il ne paraît guère possible de faire confiance aux acteurs, grandes firmes spécialement, qui sont à l'origine de ces entités. Qui serait en effet responsable des agissements de ces IA s'ils devenaient problématiques ? Personne selon toute probabilité, et en tout cas pas la puissance publique, puisqu'elle a largement disparu. Pour l'heure, dans notre début de XXIe siècle, cette dernière est encore relativement présente, mais le chantier de la régulation de cette variété avancée d'algorithmes est à peine ouvert.

On notera que, dans cette même trilogie de William Gibson, le contrôle est exercé par une police et non une agence de régulation. Autrement dit, c'est une fonction de sécurité qui est assurée par une force spécialisée, et non une branche gouvernementale. Ce qui veut dire aussi que les relations avec les acteurs soumis à juridiction sont d'une nature particulière, spécialement dans le type de confiance qui prévaut, en l'occurrence une confiance limitée puisqu'elle n'est pas accordée *a priori*. Cette police semble avoir une conception flexible, voire extensive, de sa juridiction, qu'elle adapte en effet en fonction de la situation. Case en vit la confirmation lorsqu'il est interpellé :
« *Les mecs, est-ce que vous avez réellement la moindre juridiction dans le coin ?*
[…]
Les situations d'ambiguïté, ça nous connaît. Les traités aux termes desquels notre section du Registre opère nous laissent une grande marge de manœuvre. Et nous savons la créer nous-mêmes, lorsque la situation l'exige. »[63]

Par contraste avec la tendance actuelle, l'existence de cette police laisse aussi entendre qu'il ne faut guère compter sur une autorégulation de la part des firmes. En l'absence de contraintes régulatrices, des firmes comme Tessier-Ashpool ne sont pas du genre à considérer qu'elles ont une responsabilité morale. À ce stade, il ne peut plus être question de proposer un code d'éthique aux firmes impliquées, puisqu'une telle solution, en train de se répandre sous des

[61] Rudy Rucker, *Software*, traduit de l'anglais par Nathalie Serval, Paris : Opta, coll. « Galaxie-bis », 1986 (*Software*, New York : Ace Books, 1982).
[62] Anna McFarlane, « AI and Cyberpunk Networks », in Stephen Cave, Kanta Dihal, Sarah Dillon (eds), *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*, Oxford : Oxford University Press, 2020.
[63] *Neuromancien*, op. cit., p. 192.

formats multiples[64], risquerait de paraître bien naïve. Comment pratiquement vérifier l'application et le respect de ce type de code de toute manière, de fait face à des firmes opaques et surpuissantes ? Grâce au secours de ses IA, la famille Tessier-Ashpool, profitant de régulières périodes de cryogénie, a même l'extravagante aspiration de rendre l'entreprise quasiment immortelle.

L'imaginaire esquissé, avec ses puissances économiques exerçant une domination écrasante, ne ménage guère de place à la possibilité de trouver des firmes vertueuses dans ce type de domaine technologique. De fait, quand on analyse leurs structures de gouvernance, comme le fait Alan Dignam, on remarque que les grandes sociétés du secteur de l'IA sont à tendance autocratique dans leur gestion et guère soucieuses de rendre des comptes[65].

Au surplus, la question n'est pas seulement celle des règles, mais aussi celle de leur mise en application. Qui a le pouvoir de corriger ou d'amener à corriger ? Les firmes accepteraient-elles de bon gré que des observateurs viennent s'immiscer dans ce que font leurs centres de recherche ? Lorsque ces entreprises développent leurs projets, c'est sans considérer qu'elles ont des comptes à rendre. La lecture de *Neuromancien* laisse imaginer que Tessier-Ashpool avait à sa disposition des chercheurs avec un cahier des charges correspondant aux objectifs de la firme et sans guère d'interférences extérieures.

S'il y a tentative de régulation sur un territoire, les firmes peuvent du reste avoir la forte tentation de poursuivre les développements de leur technologie ailleurs. Le cyberpunk a mis fréquemment en scène des espaces où les entreprises parviennent à créer des conditions privilégiées pour la protection de leurs intérêts, comme dans *Les fleurs du vide* de Michael Swanwick : « *De toute façon, mes souvenirs sont tous enregistrés et disponibles pour la prochaine IALI [Intelligence Artificielle Limitée Interactive] sur la liste. Alors je jouis d'une espèce d'immortalité à la chaîne. Quoique ce ne soit pas terriblement légal. Si je n'étais pas bien installée dans une zone commerciale corporative, ils m'auraient fait effacer.* »[66] La Suisse a fini par laisser dans l'imaginaire collectif une image qui peut expliquer que, dans *Neuromancien*, l'IA Muetdhiver y soit enregistrée : « *Berne. Elle a la nationalité suisse « restreinte », sous l'égide de l'équivalent de notre loi de 53. Construite pour la Tessier-Ashpool. Ils possèdent l'unité centrale et le logiciel d'origine.* »[67] Ce genre de localisation opportuniste, dont on pressent qu'elle n'est pas forcément réservée à la fiction, est un facteur qui ajoute à l'image d'opacité dans laquelle sont développées les IA.

La description de la façon dont la police de Turing travaille sous-entend implicitement que des critères ont été établis pour mesurer les capacités des IA, spécialement dans les moments où des seuils problématiques semblent pouvoir être franchis. Reste en effet à savoir comment repérer une IA devenant trop « intelligente » et s'il est possible de trouver un accord sur les risques que cela pourrait représenter. *Neuromancien* peut être lu comme une forme de crainte de ne pas parvenir à se prémunir contre des tendances malfaisantes chez les entités créées ou, avec moins d'anthropomorphisme, contre leurs capacités à engendrer des dommages. Dans le roman, la police de Turing échoue puisque ses agents se font tuer par Muetdhiver.

[64] Thilo Hagendorff, « The ethics of AI ethics: An evaluation of guidelines », *Minds and Machines*, 30 (1), 2020, p. 99-120.
[65] Alan Dignam, « Artificial intelligence, tech corporate governance and the public interest regulatory response », *Cambridge Journal of Regions, Economy and Society*, 13 (1), 2020, p. 37–54.
[66] Michael Swanwick, *Les fleurs du vide* (*Vacuum Flowers*, New York : Arbor House, 1987), traduit de l'anglais par Jean Bonnefoy, Paris : Denoël, 1988, p. 305.
[67] *Neuromancien*, op. cit., p. 89.

Pour une IA, la discrétion complète pourrait être une autre manière d'échapper à la surveillance, typiquement en se montrant moins intelligente qu'elle ne l'est. Dans *Le fleuve des dieux*, roman de Ian McDonald qui n'est pas sans partager quelques traits avec le cyberpunk, même s'il est plus tardif et plus exotique, les intelligences artificielles ont été classées en trois niveaux et celles du niveau le plus élevé (la « Génération Trois ») sont non seulement interdites, mais également pourchassées par un service policier spécialisé. Comme le fait remarquer un des personnages : « *Et qu'est-ce que tout cela prouve ? Juste une chose sur la nature du test de Turing en tant que test, et sur le danger de se fier à une information minimale. N'importe quelle aeai assez intelligente pour réussir au test de Turing l'est suffisamment pour savoir de quelle manière le rater.* »[68]

Un des enjeux majeurs va être d'avoir une compréhension des « décisions » susceptibles d'être prises par ce type d'artefact. Dans quelle mesure son fonctionnement et ses résultats restent-ils encore prédictibles ? Certains spécialistes vont jusqu'à affirmer qu'il devient impossible de prédire les actions qu'un système « intelligent » va entreprendre pour atteindre ses objectifs[69].

Les chants des IA au fond des réseaux de Jean-Marc Ligny[70] en fournit une version poussée loin, mais de manière plutôt bénéfique. MACNO, au départ « une pure intelligence logique, une quintessence d'IA »[71], prend son autonomie et entreprend de mettre l'humanité sur un autre chemin, moins soumis aux technologies. L'entité prend le contrôle de toute une série d'objets connectés, dont elle se sert pour arriver à ses fins, notamment en essayant de semer le doute chez les humains. Une autre manière de mettre en scène un débordement généralisé, mais cette fois pour le meilleur…

Conclusion

Dans le cyberpunk perçait déjà un trouble face aux intelligences artificielles et aux capacités que pourrait atteindre cette forme inédite d'altérité pour les humains. Comme un assemblage d'expériences de pensée, la mise en scène fictionnelle introduit l'éventualité qu'il puisse y avoir un seuil à partir duquel il devient difficile pour ces mêmes humains de savoir à quoi ils ont affaire. À l'instar de Muetdhiver dans *Neuromancien*, l'entité créée tendrait à adopter des fonctionnements qui lui sont propres et qui n'ont pas été anticipés au moment de la conception. Le doute était ainsi déjà installé sur les prétentions à garder un quelconque contrôle ou à garantir des formes d'encadrement efficaces.

La fiction a aussi l'avantage de remettre ces artefacts dans des contextes imaginables. Le paysage institutionnel dans lequel se développent de telles technologies est important. Qui peut avoir une appréciation des développements ? Comment ? En fonction des cadres institutionnels, qui est capable d'analyser et de comprendre les enjeux dans toute leur technicité et sur la gamme des conséquences possibles ? Y a-t-il des instances capables de déployer des formes d'évaluation d'impact ? Le type d'environnement sociopolitique posé comme cadre dans le cyberpunk invite à s'interroger sur ce qui peut se passer quand les institutions étatiques passent en retrait ou se font déborder par d'autres acteurs.

---

[68] Ian McDonald, *Le fleuve des dieux*, Paris : Folio SF, 2013, p. 56 (*River of Gods*, London : Simon & Schuster, 2004).
[69] Roman V. Yampolskiy, « Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent », *Journal of Artificial Intelligence and Consciousness*, 7 (1), 2020, p. 109-118.
[70] Jean-Marc Ligny, *Les chants des IA au fond des réseaux*, Paris : Baleine, 1999.
[71] Ibid., p. 106.

Le cyberpunk aidait à entrevoir en quoi les enjeux à l'horizon allaient être différents. L'arrivée de machines « intelligentes » dans le monde social n'a rien de neutre. L'intelligence artificielle n'est pas qu'un assemblage de code et d'algorithmes : derrière est présent tout un agencement sociotechnique. Aujourd'hui aussi, le monde dont émergent les IA est pour une très large part un monde fermé et opaque. La majeure partie des développements et des leviers de maîtrise dépendent de firmes puissantes, plutôt animées par des logiques de profit, ou correspondent à des intérêts étatiques (dans le domaine sécuritaire, typiquement). Certes, entre l'anticipation fictionnelle et la réalité, les utilisations peuvent s'avérer différentes. Dans le cyberpunk, les IA étaient notamment utilisées pour protéger des banques de données ou, comme chez Tessier-Ashpool, pour assurer des tâches de gestion. Dans le monde actuel, elles tendent davantage à servir à analyser des stocks de données de plus en plus massifs au sein desquels il s'agit de dégager des informations pertinentes et de les rendre valorisables. Les modèles d'affaires de ces firmes ont probablement plus d'influence sur le rôle donné aux IA que les structures de « gouvernance » envisagées pour elles. Comme le rappelle Ryan Calo, les mots sont importants en cette matière comme dans d'autres, et il n'est pas anodin que les termes « éthique » et « gouvernance » soient poussés dans les discussions, notamment parce que les perspectives correspondantes finissent presque par relativiser le rôle de l'intervention gouvernementale ou de la pression citoyenne dans l'encadrement de ce domaine[72].

Les développements en matière d'intelligence artificielle sont d'autant plus difficiles à connaître que l'ouverture et la transparence ne sont pas des caractéristiques majeures du secteur. Le niveau de technicité des dispositifs élaborés risque de les rendre inaccessibles à la compréhension de la majorité de la population, pour qui elles seraient finalement comme des « boîtes noires », à cause justement de l'opacité de leur fonctionnement interne. Même les audacieux pirates informatiques de la Trilogie de la Conurb ne sont pas à l'aise face à ces entités : ils ont appris à s'en méfier et à se tenir à distance. Les inspecter attentivement paraît presque hors de portée.

La transparence n'est pas un résultat automatique et spontané, comme le confirment les demandes dans cette direction[73]. Faut-il que le code et ses modalités d'élaboration soient ouverts (en « open source ») ? Cette « ouverture » ou facilité d'accès peut rassurer sans que la compréhension soit facilitée pour la très grande majorité. Une transparence complète sous-entend en outre une forme de confiance en espérant que les connaissances ainsi disponibles ne soient pas ensuite récupérées pour des usages plus problématiques[74]. Cependant, il n'est pas sûr que le concept de confiance fasse complètement sens s'agissant d'« intelligences artificielles »[75], en particulier parce qu'il est difficile de savoir dans quelle mesure elles peuvent être tenues pour responsables de leurs actions, questionnement qui ne devrait d'ailleurs pas dédouaner ceux qui développent ces systèmes de leur responsabilité. Ce type de technologie peut être fiable, mais pas forcément digne de confiance (c'est aussi la différence entre reliable et trustworthy en anglais). Et que faire quand on s'aperçoit qu'une IA peut mentir, comme Muetdhiver dans *Neuromancien* ?

Ce qui est intéressant dans ces fictions n'est pas la vision d'une IA qui peut paraître fantasmatique. C'est le contexte social dans lequel elle est insérée et qui fonctionne comme un

[72] Ryan Calo, « How We Talk About AI (and Why It Matters) », Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.
[73] Miriam C. Buiten, « Towards intelligent regulation of artificial intelligence », *European Journal of Risk Regulation*, 10 (1), 2019, p. 41-59.
[74] Thilo Hagendorff, « Forbidden knowledge in machine learning reflections on the limits of research and publication », *AI & Society*, 36 (3), 2021, p. 767–781.
[75] Mark Ryan, « In AI We Trust: Ethics, Artificial Intelligence, and Reliability », *Science and Engineering Ethics*, vol. 26, 2020, p. 2749–2767.

rappel de la nécessité de remettre toute technologie dans ses conditions de développement et de déploiement. Cet ancrage est aussi une leçon potentielle : si contrôle il doit y avoir, celui-ci ne serait pas à envisager simplement par rapport à la technologie elle-même, mais par rapport au système sociotechnique dont elle émane. La question ne serait alors pas tant la « gouvernance de l'intelligence artificielle » que celle du gouvernement de ce système sociotechnique expansif, en y incluant une prise en compte des structures de pouvoir et des garanties de type démocratique (information, débat public, etc.).

Typiquement, dans la Trilogie de la Conurb, la puissance des firmes ne permet d'intervenir qu'en aval de la conception des IA, et guère en amont. Comme si était anticipé toute la difficulté d'incorporer des considérations « éthiques » au moment de la rédaction des lignes de code… Dans l'imaginaire du cyberpunk, il aurait été difficile d'accorder crédit à l'idée d'« ethics by design » qui a été promue plus récemment et qui vise à ce qu'un certain nombre de valeurs jugées souhaitables soient intégrées dès la conception[76].

Avec ce qu'ajoute cet imaginaire, les métaphores inspirées de Frankenstein ne sont plus adaptées : le « monstre » du Dr Frankenstein ne pouvait pas orienter les humains à leur insu ; il n'était qu'une créature perdue dans un monde qui n'était pas fait pour lui. Les intelligences artificielles ont en revanche un monde numérique prêt à les accueillir. Pour les auteurs du cyberpunk, il semblait encore difficile cependant de complètement concevoir que des algorithmes complexes puissent un jour presque décider à la place des humains de ce qu'ils pourraient voir sur ce qui allait devenir l'équivalent du cyberespace, à savoir Internet, où les parcours s'avèrent de plus en plus subrepticement guidés.

L'inconvénient de ces représentations fictionnelles est au demeurant qu'elles ont contribué à véhiculer une vision de l'IA dénuée de matérialité, comme si elle n'était pas dépendante de ressources physiques et d'infrastructures potentiellement lourdes. Des travaux récents, comme ceux de Kate Crawford[77], essayent de remettre en avant cette dimension, de même que les systèmes de pouvoir auxquels participent les IA. À leur manière, par le détour de l'exploration fictionnelle, les récits et mises en scène du cyberpunk étaient déjà plus proches de cette deuxième question majeure, qui est aussi éminemment politique.

À la différence de ce qui était imaginé dans la Trilogie de la Conurb avec le registre de Turing, il n'y a pour l'heure pas encore eu de véritables réflexions sur les limites qu'il s'agirait de ne pas dépasser dans les capacités des entités créées. Comment les fixer en effet ? Dans *Neuromancien*, les humains s'en sortent favorablement, puisque les IA fusionnées se désintéressent de leur médiocre sort et préfèrent chercher les signes d'existence d'entités équivalentes à elles jusqu'ailleurs que sur Terre. Pas sûr que ces mêmes humains puissent toujours avoir autant de chance…

[76] Ron Iphofen & Mihalis Kritikos, « Regulating artificial intelligence and robotics: ethics by design in a digital society », *Contemporary Social Science*, 16 (2), 2021, p. 170-184.
[77] Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven : Yale University Press, 2021.