Dossier #3

Innovation éthique

Imaginaires de futurs, l'IA et l'éthique

Nicolas Minvielle

Olivier Wathelet

Nicolas Minvielle est professeur de design et de stratégie à Audencia Nantes. Olivier Wathelet, est anthropologue et fondateur de l'agence Users Matter. Ils sont tous les deux co-fondateur de Making Tomorrow, un collectif de designers, de makers, d'anthropologues, d'auteurs de science-fiction, de prospectivistes et d'économistes qui jouent avec le futur.

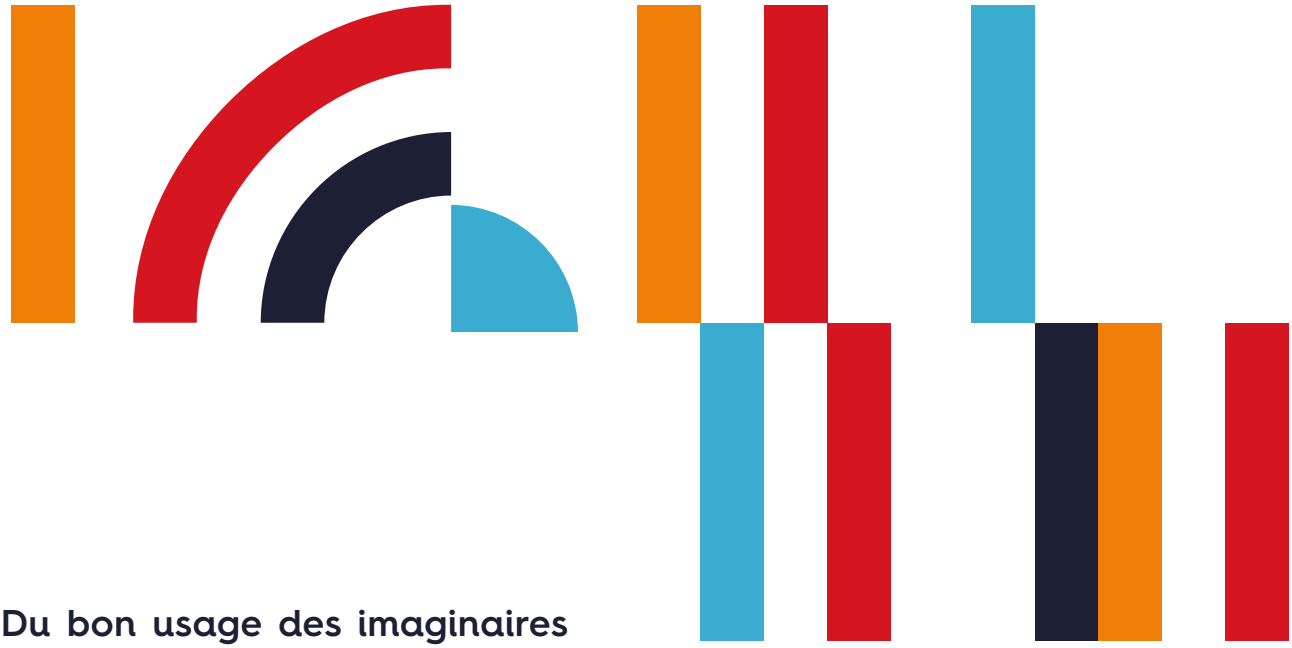
Imaginaires et innovation : stimulation ou fermeture ?

La science-fiction, et plus largement les imaginaires de futurs présents dans la culture populaire, bénéficient depuis un peu moins de 10 ans d'un intérêt grandissant dans le secteur de l'innovation. Leur capacité à s'adresser à un large public en fait des objets interpellant dont des acteurs toujours plus nombreux tentent de s'approprier la richesse et le potentiel pouvoir. S'il ne manque pas de travaux afin de démontrer le rôle des imaginaires pour tenter d'infléchir l'opinion publique, préparer un marché ou plus largement coordonner et piloter l'action collective[1], on peut se demander dans quelle mesure sont-ils également des outils adéquats pour penser et imaginer l'avenir dans une démarche de conception ? Autrement dit, quelle place donner à ces imaginaires dans toute démarche éthique en prise avec un travail d'innovation ?

Le cas des imaginaires de l'IA est intéressant pour révéler ce potentiel, dans la mesure où les nouvelles formes d'humanité sont un enjeu récurrent de la science-fiction, et que les initiatives et projections en la matière sont légions. Tant du point de vue de la culture pop (qui est une industrie), que des projections produites et manipulées par les (autres) entreprises ou les pouvoirs publics, les imaginaires se répondent de manière très homogène, tout en soulevant des débats clivants entre partie prenantes.

Le co-développement des imaginaires et des technologies, identifié sous le nom de « *loop looping* » et reposant sur la porosité réelle entre créateurs d'imaginaires à destination du divertissement ou au sein des autres entreprises[2], renforce et complexifie cette proximité entre industries. A titre d'exemple, les ingénieurs ayant participé au développement de l'assistant vocal de Microsoft (2014) soulignent l'inspiration qu'a été pour eux *Cortana*, l'assistante holographique du héros du jeu vidéo de *Halo* (2001) dont ils ont préservé le nom, à la suite d'une mobilisation des communautés de fans[3]. Plus récemment, le choix de Facebook de structurer ses projets de développement autour du concept de « *metavers* » témoigne de cette connaissance que les industries ont du potentiel mobilisateur (au sein de la R&D et auprès des clients) des imaginaires forgés et enrichis par la science-fiction.

[1] Voir la synthèse des effets économiques et politiques des récits d'avenir réalisée par le sociologue Jens Beckert dans *Imagined futures : fictional expectations and capitalist dynamics*, Harvard University Press.
[2] Voir le travail de David Kirby sur la circulation des imaginaires entre monde du cinéma et de la recherche en général, notamment dans *Lab Coats in Hollywood*, MIT Press (2011).
[3] Le jeu *Halo* a été créé par Bungie Studios, repris par 343 Industries en 2000 et détenu par Microsoft Studios. Cela souligne la porosité entre les départements fictionnels et les produits plus classiques au sein de l'entreprise.



Du bon usage des imaginaires

Au regard de ces premiers éléments, on peut raisonnablement affirmer que les œuvres de fiction et leurs déclinaisons publiques (œuvres de fan, réception grand public, présence dans d'autres supports écrits et visuels, etc.) peuvent être définies comme un laboratoire d'imaginaires de futurs qui ont tout à la fois le potentiel d'ouvrir, mais aussi de contraindre, le travail des innovateurs et de leurs « clients ». A ce double titre (prendre du recul & concevoir), une démarche d'identification et d'analyse des imaginaires d'un domaine d'activité présente de la valeur.

L'enjeu est donc de mieux comprendre ces mécanismes d'une part, et d'autre part, d'identifier des démarches capables d'en prendre appui pour améliorer notre capacité de prise de décision dans l'innovation.
Le *design fiction* n'est pas la première tentative pour capitaliser sur la richesse présumée de ces matériaux afin de stimuler l'exploration créatrice dans des démarches de conception. La plupart cherche plutôt à mettre en avant une typologie d'archétypes supposément universels au travers des œuvres[4]. Le *design fiction* propose plutôt de valoriser la capacité de mise en scène réaliste de mondes alternatifs, plausibles et à divers degrés problématiques, pour tester la pertinence de certaines réponses techniques et sociales, ainsi que stimuler l'activité de génération de solutions réalistes pour demain.

Dans notre activité au sein du collectif *Making Tomorrow* notre première démarche en ce sens consiste à collecter de très nombreux imaginaires (entre 50 et 100 par projet). Pour améliorer la capacité exploratoire, on recommande de diversifier les sources du point de vue de leur année de création (pour rendre compte d'enjeux sociaux différents), des médias supports (nous avons découvert que certaines technologies mais aussi valeurs étaient différemment représentées par type de support) et des aires culturelles d'origine (*comics*, manga, B.D. franco-belges, mais aussi *novelas* latines, romans afrofuturistes, etc.).

Un « bon » imaginaire n'est pas un film ou un roman, mais une séquence précise au sein de ces œuvres. Car c'est à l'échelle de l'événement, de l'interaction et de l'expérience, que les imaginaires deviennent concrets, critiquables et actionnables. Nommer le défi de la coexistence des IA et des humains dans le film *Blade Runner 2049* ne suffit pas, encore faut-il choisir telle séquence où cet attachement mutuel prend une forme particulière ainsi qu'interroger en profondeur de possibles scénarios, par exemple en s'appuyant sur la scène du repas pour interroger la qualité physique d'une relation avec une représentation holographique.

[4] Voir par exemple l'intéressante démarche MODIM, illustrée par Pierre Musso et ses collègues dans *Innover par les imaginaires*, éditions Manucius (2014).

Depuis le début de notre activité, nous avons ainsi travaillé avec plus de 1000 extraits. Cependant, la démarche suppose de remettre sans cesse sur le chantier ce travail pour extraire les séquences les plus pertinentes en lien avec chaque thématique, mais aussi chaque questionnement. En effet, quatre grandes catégories d'objectifs peuvent être poursuivies :

- Identifier des usages ou propositions stimulantes et créatives. C'est certainement le plus attirant mais aussi le plus difficile à mettre en place tant il faut avoir une expertise et une forte capacité d'immersion dans les imaginaires pour identifier ces « pépites ». Elles sont rarement suffisantes par elles-mêmes et c'est dans la création de nouveaux scénarios, combinant le présent et ces morceaux de futur, que des propositions nouvelles émergent.

- S'appuyer sur des mondes crédibles pour explorer la viabilité de projets est l'approche la plus fréquente. On opère ainsi à la manière d'un cinéaste qui fait vivre ses idées dans un environnement qui s'impose à lui, le contraint et ainsi développe la créativité. Le croisement entre projet et mondes illustrés par des séquences de films pertinentes ouvre la réflexion et soutient la créativité. Ce faisant, un important travail de nature prospective peut accompagner cette démarche pour ancrer les propositions dans la complexité des enjeux du présent.

- Ensuite, nous aider à prendre du recul vis-à-vis des imaginaires qui, par leur récurrence, risquent d'enfermer l'audience dans une vision évidente. Notre expérience nous a convaincu qu'il était important de conduire cette approche de manière systématique afin d'éviter de confondre probable et préférable[5]. C'est donc une approche préalable à toutes nos démarches.

- Enfin, une dernière approche est celle dite du *red-teaming*, qui consiste à jouer une situation d'adversité radicale afin de tester la capacité de réaction d'un système, d'une offre, et plus largement des organisations. Dans notre travail d'animation de la *Red Team* des armées françaises, nous avons la chance de coordonner le travail d'auteurs de science-fiction[6]. Les bénéfices sont clairement une capacité narrative riche au service d'hypothèses cohérentes mais singulières. L'enjeu est en revanche d'animer un groupe hétérogène de professionnels de la projection, chacun présentant des compétences et sensibilités singulières, mais aussi de rendre le propos crédible au regard d'un monde professionnel tout à fait spécifique. Les récits les plus efficaces développés à ce jour sont précisément ceux qui ont osé le parti-pris le plus radical, en combinant une forme de prolongement d'une tendance forte et problématique du moment avec un *deus ex machina* aux conséquences dévastatrices car aux motivations mal comprises. En ce sens, le décentrement est plus anthropologique que technologique[7].

[5] Le travail mené par Chris Noessels est assez représentatif de ce type de démarches de « nettoyage » nécessaire pour appréhender une nouvelle thématique. Voir son travail remarquable dans le domaine des IA.
[6] Placée sous l'égide de l'Agence de l'innovation de défense (AID) en coopération avec l'État-major des armées (EMA), la Direction générale de l'armement (DGA) et la Direction générale des relations internationales et de la stratégie (DGRIS), la *Red Team* offre une vision prospective afin d'anticiper les risques technologiques, économiques, sociétaux et environnementaux susceptibles d'engendrer de potentielles conflictualités à l'horizon 2030-2060. La Red Team défense est composée d'une dizaine d'auteurs et de scénaristes de science-fiction travaillant étroitement avec des experts scientifiques et militaires.
[7] Ainsi, dans le scénario Barbaresque 3.0 la pirate Alia N'Saadi s'oppose aux forces conventionnelle en faisant usage d'un détournement du NeTAM, protocole d'interface neurale Air Terre Mer. Dans un monde ou l'industrie logistique, et notamment maritime, repose prioritairement sur l'automatisation des procédures, la composante cyber devient un point faible qu'il faut à tout prix sécuriser. Le scénario explore les conséquences de cette faiblesse, et en particulier la dimension collective et en réseau de ces dispositifs neuronaux d'augmentation des compétences cognitives et sensorielles. Une des originalités du scénario est d'avoir choisi de raconter une intoxication longue et graduelle des opérateurs connectés, altérant progressivement leurs cadres cognitifs, leur mémoire, afin de rendre possible une attaque fulgurante le jour J sans que le Cheval de Troie ne soit détecté.

Les imaginaires de l'IA : un territoire prêt-à-penser ?

Pour illustrer la pertinence d'un travail sur les imaginaires, prenons donc l'exemple des IA. Que disent les imaginaires en la matière ? De prime abord, leur absence de neutralité est assez frappante. En effet, ils tendent avant tout à interroger les défis et limites de la cohabitation avec les humains. Nombreuses sont les IA tueuses, soit du fait d'une programmation volontairement létale (T*erminator*, 1984, *Robocop* manipulé par l'entreprise qui l'a fondée, 1987, *Assassination Classroom*, 2012), soit du fait de dérives d'algorithmes agissant de manière trop « rigoureuse » (*Alphaville*, 1965, *Logan's Run*, 1976, *Proteus IV*, 1977, ou quand l'IA de *Hotel Since AD2019*, 2008 poursuit durant des siècles un même programme, voir aussi la version positive de la tâche « obstinée » de *Wall-E*, 2008).

A contrario, il existe aussi une riche production d'œuvres, qui cherche à multiplier les points de vue et propose de considérer le statut de sujet des IA. Dans la série de *comics Top10* imaginée par Alan Moore (1997-2002), les « post organiques » ou « ferro-americans » sont considérés comme un groupe ethnique à part, discriminé et péjorativement nommé « clickers ». La série montre toutefois comment, par le sens de l'humour et le partage de défis communs, deux partenaires (l'un humain, l'autre robot) vont se lier d'amitié et se donner une place respective. Plusieurs œuvres testent ainsi la possibilité et les limites d'un compagnonnage poussé entre humains et IA, voire des humains « augmentés » avec leur propre IA (*Upgrade*, 2018). A l'instar du super héros *The Vision*, qui tentera en vain de rendre ordinaire sa famille, les imaginaires tendent à montrer la difficulté qu'il y a pour donner place à des compétences plus qu'humaines dans un monde qui n'est pas adapté à l'extraordinaire (voir aussi sur le lien entre ordinaire et extra-ordinaire dans le manga *Eve no Jikan*, 2010, ou encore le comics *Descender*, 2016). La frontière tend au final à être mince entre quête d'autonomie (*D.A.R.Y.L.L.*, 1985, *Johny 5*, 1986, *A.I*, 2001) et désir de rébellion contre ses « créateurs » (*Colossus. The Forbin Project*, 1970, *Westworld*, 1973, *Ex Machina*, 2014).

Au-delà de ces grandes thématiques, ce qui nous intéresse porte plus volontiers sur les techniques et interactions mises en scènes dans certaines séquences de ces œuvres. Les imaginaires explorent ainsi volontiers les dispositifs permettant de faciliter, voire de pacifier, la relation entre humains et IA. Sont décrits des commandes de sécurité (la plus fameuse étant sans nulle doute « *Klaatu barada nikto* » du film *The Day the Earth Stood Still*, 1952), des tests de détection du statut d'IA (le paradigmatique test de Voight-Khampf de *Blade Runner*, 1982), de transparence dans le processus de décision (*Forbiden Planet*, 1956, *Interstellar*, 2014) ou encore la capacité offerte aux IA de s'opposer aux ordres malveillants des « propriétaires » de la robotique (la B.D. argentine *El Humano*, 2019).

La complexité des interactions est également racontée dans des œuvres où l'attachement se noue (*Real Human*, 2012, *Her*, 2013, *Alex + Ada*, 2016, le barman de *Passengers*, 2016). Une autre voie intéressante à nos yeux est celle qui explique un formalisme peu ou pas du tout anthropomorphique, explorant des dispositifs parfois « simplistes » (*Moon*, 2009, *Fallout*, 2018), reposant sur des attributs d'interaction et non des composantes formelles.

Ce que cette rapide revue montre est que les imaginaires de l'IA offrent un terrain d'exploration très riche des enjeux de la coexistence entre humains et IA. Toutefois, la récurrence de certains thèmes – le *pattern* des IA tueuses ; la séduction de l'IA anthropomorphe – présente le risque d'enfermer les visions à la manière de *La Famille Jetson* (1962) qui serait l'archétype de la maison moderne, autonomisée. C'est pour cela que, de même que l'avenir préférable de la vie domestique n'est ni tout à fait son contraire, ni son prolongement direct, un enjeu de toute démarche éthique en la matière consiste sans nul doute à s'autonomiser de ses visions pour développer un horizon propre à soi.

Ce que le *design fiction* peut apporter aux réflexions sur l'IA

En tant que pratique issue du *design* critique (Minvielle & Wathelet 2017), le *design fiction* est aujourd'hui proposé comme une réponse à ces défis. Il s'agit de concevoir des fictions alternatives (à celles disponibles) en vue de projeter une audience dans un futur crédible, réaliste, et générer un débat quant au caractère préférable de l'avenir. Le débat n'est jamais une fin en soi, mais une étape intermédiaire vers la production d'éléments stratégiques (feuille de route, vision d'entreprise) ou opérationnels (identifier et initier des concepts innovants, définir une trajectoire de « petits pas » vers une transformation, etc.).

L'enjeu clé de cette approche réside dans sa capacité à faire évoluer des perceptions individuelles. Les œuvres produites dans ce domaine ont ainsi pour l'essentiel une portée critique, à l'instar du projet de santé autonome *Jewels* que nous avons créé pour le compte du programme Interreg GoToS3[8].

Toutefois, derrière la conviction partagée au sein de la communauté des praticiens du *design fiction* que ce type d'approche « fait changer » le regard, il manque encore des études pour en démontrer l'impact et structurer la pratique dans des directions meilleures. Il s'agit donc d'un défi de taille pour un travail réellement éthique.

Sur la base de notre expérience, et de premières expérimentations auprès de publics étudiants, nous faisons l'hypothèse que les fictions générées par cette approche ne permettent pas de produire du débat et tendent au contraire « naturellement » à renforcer les attitudes antérieures. Ce point a récemment été démontré dans une étude portant sur les robots tueurs et visant à définir si la science-fiction affecte les perceptions politiques[9]. Les auteurs ont testé un certain nombre de variables sur des Américains afin de définir dans quelle mesure des fictions négatives (*Odyssée*, 2001) ou positives (*Star Trek*) pouvaient impacter les perceptions politiques quant aux « robots tueurs ».

Deux principales conclusions nous intéressent ici. Tout d'abord le fait que les récits sombres (dystopies) ont un impact beaucoup plus important sur les changements de perception que les utopies. En termes de pratiques de *design fiction*, c'est un point clé qui doit nous amener à éviter de proposer des fictions clivantes (faire du « *black* » ou « *bright* » futur selon la terminologie employée par les sociétés de conseil) d'une part, et s'interdire d'investiguer un univers sans en dresser préalablement le contour comme nous l'avons illustré par l'emploi de sources nombreuses et variables.

Ensuite, il nous semble aujourd'hui important de valoriser dans ces approches le fait que l'impact des fictions n'est pas le même pour tous. En l'occurrence, plus on regarde de science-fiction, plus le fait de voir une fiction va avoir un impact sur des préférences individuelles. De ce point, on retrouve ici des éléments développés dans d'autres travaux et portant sur le « *fictionnal overload* »[10] qui montrent l'effet de renforcement de la consommation d'œuvres dystopiques sur des visions négatives à l'encontre d'un imaginaire.

Ces éléments plaident en faveur du développement d'outils plus fins que ceux actuellement proposés, reposant notamment, c'est notre hypothèse, sur une clarification préalable des positions et sur un mécanisme de présentation transparente des positions durant les échanges élaborés autour des fictions. Ceci, en vue de rendre possible un travail éthique partagé, nécessaire pour avancer plus sereinement dans l'évaluation des choix futurs de société d'une part, et pour innover de manière plus libre d'autre part.

En résumé, le *design fiction* est un outil qui capitalise sur nos imaginaires pour donner à voir des futurs en en produisant de nouveaux. La production actuelle desdits imaginaires est tellement massive et diversifiée que toute démarche qui se veut éthique gagne à s'en inspirer, et à en diversifier les sources. De ce point de vue, la segmentation usuelle entre dystopies et utopies n'est pas suffisante, voire a tendance à cliver les audiences.

Toutefois, en y prenant soin et à l'aide d'approches adaptées, les imaginaires populaires et ceux issus du *design fiction*, malgré leurs limites, sont des « outils à penser » incroyables. De ce point de vue, ils peuvent contribuer à améliorer le débat éthique en donnant à voir des futurs impensés ou des conséquences non souhaitées de développements technologiques, économiques ou politiques actuels.

[8] Cette œuvre et différents supports physiques ont été présentés à un parterre de plus de 300 industriels et représentants politiques comme étant un projet réel, avant d'en dévoiler le caractère fictionnel. L'enjeu était de faire réagir une audience au caractère fondé de certains choix pourtant jugés évidents dans le cadre de la programmation.

[9] Kevin L.Young & Charli Carpenter, Does science fiction affect political fact ? Yes and no: a survey experiment on "killer robots", *International Journal Quaterly Studies*, 2018, 62, 562-576

[10] Alexander H. Montgomery & Amy J. Nelson, *The rise of the futurist: the perils of predicting with futurethink*, Brookings, Novembre 2020