Éthique de la recherche en apprentissage machine

Edition provisoire, juin 2017

****

Préambule

Dans le numérique, le foisonnement et la rapidité de déploiement des usages issus de l'innovation contribuent à la complexité de l'interaction entre l'offre technologique et l'appropriation par la société, et réduit de ce fait la portée des prévisions scientifiques sur les conséquences de la recherche. Cette relative imprévisibilité des usages ne doit pas dédouaner les scientifiques, mais doit au contraire motiver la réflexion éthique et la recherche d'attitudes et de méthodes adaptées. En effet les chercheurs doivent avoir à l'esprit que leurs travaux contribuent *de facto* à transformer la société et peut-être l'Homme, comme l'ont fait beaucoup d'outils et de techniques depuis des millénaires,

même si ce processus n'est pas toujours prévisible. Ainsi si l'on ne saurait attribuer aux seuls chercheurs la responsabilité de l'impact potentiel de leurs travaux, ceux-ci doivent être conscients qu'ils sont partie prenante d'une responsabilité collective. Le monde de la recherche doit organiser en son sein et d'une manière transdisciplinaire la prise en compte de la dimension éthique, et éclairer ses choix vis-à-vis de la société en contribuant aux débats publics, afin que la science demeure un facteur de progrès et que les croyances infondées et l'irrationnel ne conduisent pas à une défiance à son égard. Dans ce contexte, les réflexions de la CERNA - dont la vocation est de se prononcer sur l'éthique de la recherche en sciences et technologies du numérique - visent à inciter et aider les chercheurs à la vigilance éthique « chemin faisant » plutôt qu'à émettre des prescriptions normatives qui seraient vite obsolètes. Elles n'envisagent que des perspectives plausibles du point de vue scientifique, afin de ne pas nourrir la confusion avec ce qui relève de la science-fiction.

Conçu dans un esprit pratique en premier lieu à l'attention des chercheurs et développeurs dans le numérique, le document aborde les questions sous l'angle des sciences et technologies. Des questions de société que le concepteur doit avoir à l'esprit sont évoquées sans être approfondies. Le présent travail n'est qu'une contribution à une réflexion qui doit être plus vaste au sein du monde de la recherche, notamment avec les sciences humaines et sociales, et au niveau de la société toute entière, comme l'évoque la conclusion.

Composition du groupe de travail Apprentissage Machine

Laurence Devillers,
Professeur Paris-Sorbonne 4, LIMSI-CNRS, CERNA, animatrice du groupe
Serge Abiteboul,
Directeur de recherche Inria, ENS-Paris, Membre de l'Académie des sciences
Danièle Bourcier,
Directrice de recherche émérite au CNRS, CERSA, CERNA
Nozha Boujemaa,
Directrice de recherche Inria
Raja Chatila,
Professeur UPMC, directeur de l'ISIR, CERNA
Gilles Dowek,
Directeur de recherche Inria, ENS-Saclay, CERNA
Max Dauchet,
Professeur émérite, Université de Lille, président de la CERNA
Alexei Grinbaum,
Chercheur CEA, IRFU/LARSIM, CERNA
Avec la collaboration de Christophe Lazaro, chercheur à l'Université de Namur, CERNA, et de Jean-Gabriel Ganascia, Professeur UPMC Paris 6, LIP6, CERNA

Personnalités auditionnées (groupe restreint)

Alexandre Allauzen,
Maître de conférences, Université Paris 11, LIMSI
Edouard Geoffrois,
Responsable de programme du département de l'Information et des sciences et technologies de la communication, ANR
Mathieu Lagrange,
Chercheur CNRS, LS2N
Arnaud Lallouet,
Ingénieur en chef, Huawei Technologies Ltd.
Olivier Teytaud,
Chargé de recherche, Inria

Personnalités auditionnées lors de la journée CERNA *Apprentissage et IA* du 13 juin 2016 à Paris[1]

Tristan Cazenave,
Professeur, Paris-Dauphine, LAMSADE
Milad Doueihi,
Chaire d'humanisme numérique, Paris-Sorbonne ; co-titulaire de la Chaire des Bernardins *L'humain au défi du numérique*
Benoît Girard,
Directeur de recherche CNRS, ISIR
Jean-Baptiste Mouret,
Chercheur Inria, Equipe Larsen

Expert relecteur

Léon Bottou,
Chercheur Facebook AI : Machine learning, artificial intelligence

Introduction

L'apprentissage automatique, aussi appelé apprentissage statistique ou apprentissage machine (*machine learning*), a récemment fait des progrès spectaculaires, popularisés en 2016 par la victoire du programme AlphaGo face au champion de Go, Lee Sedol. Ses applications sont multiples : moteurs de recherche, reconnaissance d'images et de parole, traduction automatique, agents conversationnels par exemple. Elles commencent à émerger dans des secteurs comme la santé, l'énergie, les transports, l'éducation, le commerce et la banque.

Les succès de l'apprentissage machine, champ d'études de l'intelligence artificielle (IA), s'appuient sur l'accroissement des capacités de calcul, de stockage et de traitement des données (*Big data*), et font resurgir avec une médiatisation autant excessive qu'approximative l'idée que la machine - parfois un robot - pourrait apprendre à s'affranchir de l'homme. Si cette question est hors du champ de la science actuelle, il n'en demeure pas moins qu'une réflexion éthique doit éclairer le bon usage d'algorithmes d'apprentissage, et de masses de données de plus en plus complexes et disséminées. Des initiatives en ce sens, publiques ou privées, aux niveaux national, européen ou international, voient le jour depuis 2015.

Dans ce contexte, le présent travail vise à

- Sensibiliser, donner des éléments de réflexion et des repères au « chercheur ». *Par commodité le terme « chercheur » désigne ici la ou les personnes - concepteurs, ingénieurs, développeurs, entrepreneurs - leurs communautés ou institutions.*
- Contribuer plus largement au débat sur les questions éthiques et sociétales liées au développement de l'intelligence artificielle.

afin que l'apprentissage machine soit mis en œuvre au bénéfice de la société.
Ce regard de la CERNA est donc conçu pour une double lecture : celle du spécialiste et celle de toute personne intéressée, décideur ou simple citoyen.

La partie I introduit des notions de base de l'apprentissage et les illustre à travers la méthode particulière des réseaux multicouches et de l'apprentissage profond. La partie II énumère des cas d'usage déjà répandus ou appelés à le devenir. Ces deux parties fournissent un support technologique pour les réflexions éthiques et s'adressent particulièrement aux non spécialistes. La partie III présente les questions éthiques générales liées aux systèmes numériques et pointe les spécificités liées à l'apprentissage.

1http://cerna-ethics-allistene.org/journ%C3%A9e+apprentissage/

La partie IV analyse ces questions éthiques et émet des préconisations à l'attention des scientifiques et des communautés qui conçoivent et développent les systèmes apprenants. Ces préconisations sont des points d'attention et de vigilance pour susciter la réflexion éthique individuelle et collective, elles ne sauraient être des « recettes ». Elles sont articulées autour de six questions :

1. Quelles sont les données sélectionnées/utilisées à partir desquelles la machine apprend ?
2. Peut-on s'assurer que la machine effectuera uniquement les tâches pour lesquelles elle a été conçue ?
3. Comment peut-on évaluer un système qui apprend ?
4. Quelles décisions peut-on déléguer, ou non, à un système apprenant ?
5. Quelle information doit-on donner aux utilisateurs sur les capacités des systèmes apprenants ?
6. Qui est responsable en cas de dysfonctionnement de la machine : le concepteur, le propriétaire des données, le propriétaire du système, son utilisateur ou peut-être le système lui-même ?

Les initiatives évoquées dans la partie V illustrent l'actualité des questionnements éthiques liés aux développements de l'apprentissage machine et plus généralement de l'intelligence artificielle. La partie VI conclut par des préconisations générales à l'attention des opérateurs scientifiques et des décideurs de la société.

I. Qu'est-ce que l'apprentissage machine ?

Construire des systèmes capables de fonctions de perception, d'apprentissage, d'abstraction et de raisonnement est un des buts des chercheurs en intelligence artificielle. Pour ce faire, les algorithmes d'apprentissage utilisent différentes méthodes statistiques en se basant sur des données d'apprentissage, par exemple pour construire des règles de déduction, des arbres de décision ou pour paramétrer des réseaux de neurones, puis les appliquent à de nouvelles données.

Prédire un phénomène à partir d'observations passées présuppose un mécanisme causal. Expliquer ce mécanisme n'est pas toujours facile. L'apprentissage machine est une approche statistique permettant de découvrir des corrélations significatives dans une masse importante de données pour construire un modèle prédictif quand il est difficile de construire un modèle explicatif. La reconnaissance de l'écriture manuscrite est un exemple de problème difficile pour une machine. Pour reconnaître une lettre, ou un chiffre, certains algorithmes utilisent des règles préétablies, mais d'autres algorithmes « apprennent » à reconnaître les lettres de l'alphabet, à partir d'un grand nombre d'exemples. Ces algorithmes, qui utilisent des données pour apprendre à résoudre un problème, sont appelés « algorithmes d'apprentissage machine ». Ils se développent dans de nombreux champs d'application, comme la finance, le transport, la santé, le bien-être ou encore l'art .

Par exemple, dans le domaine du transport, des systèmes obtenus par apprentissage machine sont utilisés pour reconnaître visuellement l'environnement routier des voitures autonomes. Dans un tout autre domaine, la reconnaissance faciale popularisée par GoogleFace et Facebook est utilisé dans les réseaux sociaux pour identifier des personnes dans des photos. Dans le domaine du jeu, le système *Deep Blue*[2]

2Deep Blue est un superordinateur spécialisé dans le jeu d'échecs développé par IBM au début des années 1990.

d'IBM a gagné contre le champion du monde d'échecs dès 1997. En 2011, *Watson*[3] d'IBM a participé à trois manches du jeu télévisé *Jeopardy*, au terme desquels il a remporté la partie. A*lphago*[4] de Google *DeepMind* a vaincu en 2016 au jeu de go un des meilleurs joueurs mondiaux, Lee Sedol.

Actuellement, un champ de recherche émerge afin d'améliorer l'explicabilité et la transparence des systèmes d'apprentissage ainsi que leur adaptation en contexte et l'adéquation de l'apprentissage à ce qu'en attend l'humain. Ainsi il ne s'agit plus seulement de construire des modèles par apprentissage machine sans comprendre mais bien d'essayer de les expliquer.

I.1 Les différents types d'algorithmes d'apprentissage machine

Les algorithmes d'apprentissage machine sont nombreux et divers. Ils peuvent se classer en trois grandes catégories selon leur mode d'apprentissage *supervisé*, *non supervisé* et *par renforcement*.

En apprentissage supervisé, les données utilisées doivent être annotées préalablement par des « experts ». Par exemple, pour construire un système prédictif de reconnaissance de lettres dans des images, les experts indiquent les images de l'ensemble de données qui représentent des « a », des « b », etc. Lors d'une première phase, dite d'*apprentissage*, la machine construit un « modèle » des données annotées, qui peut être un ensemble de règles, un arbre de décision, un ensemble de matrices comme dans les réseaux de neurones, etc. Ce modèle est ensuite utilisé dans une seconde phase, dite de *reconnaissance*, dans laquelle par exemple l'algorithme reconnaît une lettre dans une nouvelle image. Les machines à vecteurs de support[5] (*support vector machine*, SVM), ou les réseaux de neurones de type perceptron multicouches à rétropropagation du gradient[6] sont des exemples d'algorithmes d'apprentissage machine supervisé.

*En apprentissage machine non supervisé***,** aucun expert n'est requis pour annoter les données. L'algorithme découvre par lui-même la structure des données, en classant ces données en groupes homogènes. Les k-moyennes[7] (*k-means*, méthode de partitionnement de données ou *clustering*) et les réseaux de neurones de type carte de Kohonen[8] (méthode de réduction de dimensions) sont des exemples d'algorithmes d'apprentissage machine non supervisé.

En apprentissage par renforcement, le but est d'apprendre, à partir d'expériences, ce qu'il convient de faire en différentes situations, de façon à optimiser une récompense quantitative au cours du temps. L'algorithme procède par essais et erreurs, chaque erreur l'amenant à améliorer sa performance dans la résolution du problème. Le rôle des experts

3Watson est un programme informatique d'intelligence artificielle conçu par IBM dans le but de répondre à des questions formulées en langage naturel.

4AlphaGo est un programme informatique conçu pour jouer au jeu de go, développé par l'entreprise britannique Google DeepMind.

5 Boser Bernhard E., Guyon, Isabelle M., Vapnik, Vladimir N. , "A training algorithm for optimal margin classifiers" COLT'92, pp.144-152

6Rumelhart, David E. Hinton, Geoffrey E., Williams, Ronald J. (8 October 1986). "Learning representations by back-propagating errors". *Nature*. 323 (6088) : 533–536

7MacQueen, J. B. Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.* 1. University of California Press. pp. 281–297 (1967)

8Kohonen, Teuvo (1982). "Self-Organized Formation of Topologically Correct Feature Maps". *Biological Cybernetics. 4*3 (1) : 59–69

se limite ici à définir les critères de succès de l'algorithme. Les algorithmes TD-learning[9] et Q-learning[10] sont des exemples d'algorithmes d'apprentissage par renforcement.

Certaines méthodes enfin sont intermédiaires, tel l'apprentissage semi-supervisé, qui laisse parfois la place à l'intervention humaine mais soulève des contraintes de temps réel. Plusieurs méthodes d'apprentissage sont en outre souvent combinées dans un même système.

I.2 Un exemple : les réseaux de neurones multicouches

Les réseaux de neurones multicouches sont entraînés avec des algorithmes d'apprentissage machine. Leur conception est à l'origine très schématiquement inspirée du fonctionnement des neurones biologiques. Ils utilisent un concept de neurone formel issu d'une analogie avec les neurones du cerveau. Un neurone formel, Fig. 1, comme un neurone naturel, a de multiples valeurs d'entrées qui déterminent une unique valeur de sortie, transmise comme entrée à d'autres neurones.

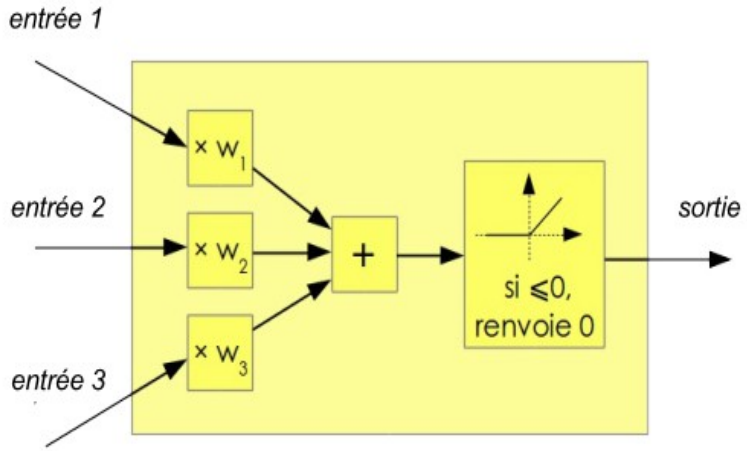

Figure 1 : Le neurone formel de McCulloch et Pitts est un modèle mathématique très simple dérivé d'une analyse de la réalité biologique (Michel Blancard[11])

Cette sortie peut être une simple combinaison linéaire des entrées

$$y = w_1\ entrée_1 + w_2\ entrée_2 + \ldots + w_n\ entrée_n$$

ou la composée d'une telle combinaison linéaire avec une fonction d'activation (fonction à seuil, fonction sigmoïde, etc.). Les poids synaptiques w_1, w_2… de chaque neurone sont déterminés itérativement au cours de la phase d'apprentissage sur les données annotées. Cette capacité des neurones à changer ces poids au cours du temps est appelée « plasticité cérébrale »[12].

9 Sutton, R.S., 1988, Learning to Predict by the Method of Temporal Differences, *Machine Learning*, 3, pp. 9-44

10 Watkins, C.J.C.H. & Dayan, P., *Q-learning*, Mach Learn (1992) 8 : 279

11 Data-scientist au sein de l'équipe de l'Administrateur Général des Données, https://agd.data.gouv.fr/

12 Hebb, D.O., The Organization of Behavior, New-York, Wiley and Sons, 1949.

Le premier réseau, le Perceptron de Rosenblatt, datant des années 50 n'a qu'une couche. Certains systèmes, tel le Perceptron multicouche, empilent plusieurs couches de neurones formels afin de pouvoir reconnaître une forme comme par exemple une image - sans pour autant inférer les concepts humains ni la logique les articulant. Une couche peut comporter des milliers de neurones, et donc des millions de paramètres. Entre la couche d'entrée et la couche de sortie, le réseau peut comporter plusieurs dizaines de couches dites cachées. Les systèmes d'apprentissage profond (*deep learning*)[13] sont des réseaux de neurones comprenant un grand nombre de couches.

La phase d'apprentissage détermine les valeurs des poids synaptiques à partir d'un échantillon de données de très grande taille (jusqu'à des millions). Dans l'algorithme d'apprentissage supervisé de *rétro-propagation du gradient*[14], l'écart entre les sorties attendues et les sorties constatées est diminué pas à pas (*descente de gradient*) en modifiant les paramètres, de la sortie vers les premières couches (*rétro-propagation*), jusqu'à obtenir un minimum local (le minimum absolu est difficilement atteignable). La valeur initiale des poids synaptiques est parfois tirée au hasard, elle peut aussi être déterminée par un algorithme d'apprentissage non supervisé.

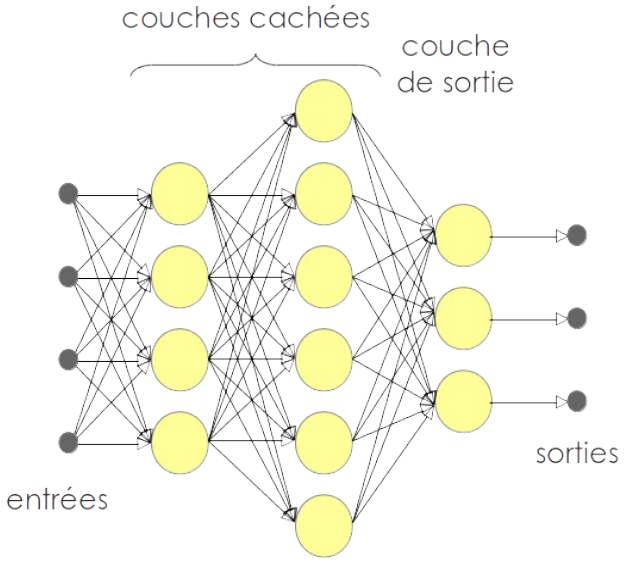

Figure 2 : Le perceptron multicouche avec ici une couche cachée (Michel Blancard)

Concevoir un réseau profond qui apprend à classifier de manière satisfaisante demande beaucoup d'expertise et d'ingénierie, et le terme « satisfaisant » est à prendre en un sens empirique, où pour des données de situations réelles les résultats sont conformes à ce qui est attendu. Comme le souligne Yann LeCun, l'apprentissage profond exploite la structuration modulaire des données du monde réel[15]. Son succès réside dans sa grande capacité d'apprentissage sans qu'il soit besoin d'expliciter un modèle des données. Le

13Y. LeCun, Y. Bengio et G. Hinton, (2015) « Deep learning », *Nature*, vol. 521, no 7553, 2015, p. 436–444 (PMID 26017442, DOI 10.1038/nature14539

14Rumelhart, David E.,Hinton, Geoffrey E., Williams, Ronald J. (8 October 1986). "Learning representations by back-propagating errors". *Nature*, vol. 323 n° 6088, p. 533–536

15Yann LeCun, Chaire Informatique et Sciences numériques Collège de France/Inria, 2015-2016, https://www.college-de-france.fr/site/yann-lecun/

cadre mathématique de la descente de gradient justifie ces méthodes, mais n'assure pas la réussite de l'apprentissage (des théorèmes de convergence n'existent que dans des cas très simples) et ces algorithmes demandent un grand nombre d'itérations pour converger empiriquement vers une solution acceptable. Les succès récents de ces méthodes doivent beaucoup à l'augmentation de la puissance de calcul des machines et au grand nombre de données disponibles.

L'architecture (type de neurones, choix des connexions) doit être adaptée au domaine d'application. Il existe différentes architectures de systèmes de *deep learning*, par exemple les réseaux récurrents ou convolutifs et des approches complexes combinant plusieurs systèmes de *deep learning*.

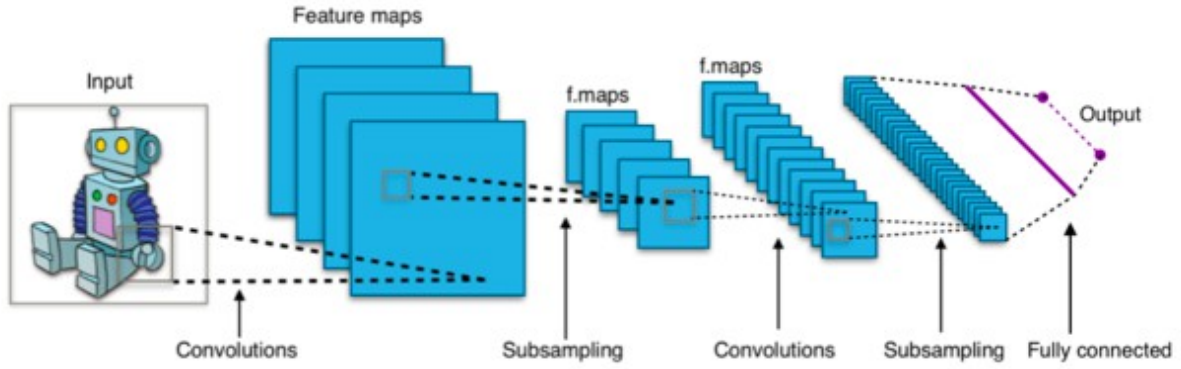

Figure 3 : Architecture standard d'un réseau convolutif[16]

Par exemple, le système AlphaGo développé par Google Deep Mind est une combinaison de recherche arborescente Monte Carlo et de Deep Learning. Un réseau profond est entraîné sur des parties de joueurs humains de haut niveau afin de prévoir leurs coups. Ce réseau atteint un niveau de 3e Dan et améliore son jeu à l'aide d'apprentissage par renforcement en jouant 30 millions de parties contre lui même. Il utilise alors les résultats de ces parties pour entraîner un autre réseau qui apprend à évaluer des positions. La recherche arborescente Monte Carlo utilise le premier réseau pour sélectionner les coups intéressants et le second réseau pour évaluer les positions. Serge Abiteboul et Tristan Cazenave précisent la typologie des réseaux profonds d'AlphaGo et le principe de Monte-Carlo dans le bulletin de la SIF 2016[17] : « *Les réseaux utilisés pour AlphaGo sont par exemple composés de 13 couches et de 128 à 256 plans de caractéristiques. Pour les spécialistes : ils sont « convolutionnels », avec des filtres de taille 3 × 3, et utilisent le langage Torch, basé sur le langage Lua* (…) *Le principe de la recherche Monte-Carlo est de faire des statistiques sur les coups possibles à partir de parties jouées aléatoirement. En fait, les parties ne sont pas complètement aléatoires et décident des coups avec des probabilités qui dépendent d'une forme, du contexte du coup* (disposition des pierres sur le goban)*. Tous les états rencontrés lors des parties aléatoires sont mémorisés et les statistiques sur les coups joués dans les états sont aussi mémorisées. Cela permet lorsqu'on revient sur un état déjà visité de choisir les coups qui ont les meilleures statistiques. AlphaGo combine l'apprentissage profond avec la recherche Monte-Carlo de deux façons. Tout d'abord il utilise le premier réseau, qui prévoit les coups, pour essayer en premier ces coups lors des parties aléatoires. Ensuite il utilise le second réseau, qui évalue les positions, pour corriger les statistiques qui proviennent des parties aléatoires*. »

16" Convolutional Neural Networks (LeNet) – DeepLearning 0.1 documentation " sur *DeepLearning 0.1*, LISA Lab

17"Go : Une belle victoire... des informaticiens !", Serge Abiteboul, Tristan Cazenave. Bulletin n° 8 de la SIF 2016

Le champ de recherche de l'apprentissage profond représente une part importante des articles parus dans les revues de référence en apprentissage automatique. Yann LeCun a donné un cours sur ce sujet dans le cadre de la chaire annuelle « Informatique et sciences numériques » du Collège de France.

Des plates-formes comme Aifree mettent à disposition une panoplie de réseaux permettant aux non spécialistes d'expérimenter ou de développer des applications de l'apprentissage machine. D'autres initiatives visent à mettre les plate-formes d'apprentissage à la portée de tous. Google rend accessible sa plate-forme *DeepDream* d'expérimentation du Deep Learning, ainsi que son outil open source d'apprentissage automatique, *TensorFlow.* Facebook fait de même avec son serveur Big Sur pour exécuter de grands réseaux neuronaux d'apprentissage. Par ailleurs, *Open IA* a été créée récemment avec d'importants financements privés (1 milliard de dollars), notamment d'Elon Musk, de Peter Thiel et de Reid Hoffman qui déclarent « *Notre objectif est de faire progresser l'intelligence numérique de la manière qui est la plus susceptible de bénéficier à l'humanité dans son ensemble, sans contrainte à générer un rendement financier* ». Les initiatives de ce type permettent une acculturation et un développement collaboratif des outils d'apprentissage, et l'approche ouverte vise à leur contrôle collectif. Cependant, elles soulèvent la question d'une prolifération d'applications mal maîtrisées et non sécurisées que pourraient développer des particuliers et certaines start-up.

II. Exemples d'applications de l'apprentissage machine

« *Il est très probable qu'à l'heure où vous lisez ces lignes, vous aurez utilisé le résultat d'algorithmes d'apprentissage automatique plusieurs fois aujourd'hui : votre réseau social favori vous a peut-être proposé de nouveaux amis et le moteur de recherche a jugé certaines pages pertinentes pour vous mais pas pour votre voisin. Vous avez dicté un message sur votre téléphone portable, utilisé un logiciel de reconnaissance optique de caractères, lu un article qui vous a été proposé spécifiquement en fonction de vos préférences et qui a peut-être été traduit automatiquement* ». (Colin de la Higuera, Blog *Binaire* du journal *Le Monde,* 23 juin 2015)

De fait, de nombreux agents artificiels utilisent des modules d'apprentissage machine. Ces agents sont des agents logiciels comme les agents conversationnels ou des agents physiques comme les robots ou les voitures autonomes. Ils ont une plus ou moins grande autonomie et peuvent apparaître comme des acteurs sociaux, avec les capacités d'interaction voire de simulation d'émotions et de décision d'actions.

II.1 Les recommandations personnalisées

Les traces que nous laissons à travers nos consultations sur Internet et à travers les objets auxquels nous sommes connectés sont exploitées par des algorithmes d'apprentissage afin de mieux cerner nos préférences de consommation, notre mode de vie et nos opinions. Par rapport aux simples statistiques, ces algorithmes ont – ou peuvent avoir – la capacité de fournir des prescriptions individuelles. Ainsi lorsque nous naviguons sur Internet et achetons des gadgets connectés, nous ne pensons pas que notre sillage numérique peut entraîner des algorithmes à nous catégoriser pour conditionner le montant de nos primes d'assurance ou conduire à des prescriptions d'hygiène de vie. La conformité, la transparence, la loyauté, l'équité des algorithmes d'apprentissage sous-jacents apparaissent ici comme une exigence.

II.2 Les agents conversationnels (bots)

Les *chatbots*, ou *bots*, sont des agents de traitement automatique de conversation en langage naturel. Ils sont de plus en plus présents en tant qu'assistants personnels ou interlocuteurs dans des transactions commerciales réalisées en ligne sur des plates-formes informatiques. Il arrive aussi qu'ils soient majoritaires dans des *chats* avec les humains. La prolifération massive de ces conversations, sans hiérarchie ou distinction claire entre l'humain et la machine, pourrait à terme influer sur le corpus des textes disponibles en ligne. D'autre part le comportement des *bots* est conditionné par les données d'apprentissage. Des *bots* apprenants sont expérimentés début 2017 par le système de soins britannique (NHS), non seulement pour désengorger les services face aux appels, mais aussi dans l'espoir que le couplage de ces *bots* aux très grandes bases de données médicales améliorera le conseil. Mais les *bots* peuvent aussi être entraînés ou utilisés à des fins pernicieuses. Ils sont d'ores et déjà utilisés à des fins d'influence, comme simples prescripteurs commerciaux ou dans le champ politique en vue d'élections. En avril 2016, le *chatbot* Tay de Microsoft, qui avait la capacité d'apprendre en continu à partir de ses interactions avec les internautes avait appris en 24h à tenir des propos racistes[18]. Tay a été rapidement retiré par Microsoft.

Les deux exemples suivants, traités dans l'avis de la CERNA *Éthique de la recherche en robotique* de la CERNA (2014)[19], sont ici brièvement évoqués du point de vue de l'apprentissage machine.

II.3 Les véhicules autonomes

Chaque accident impliquant un véhicule totalement ou partiellement autonome est abondamment commenté dans les médias[20]. Pourtant, sur 10 millions d'accidents par an aux États-Unis, 9,5 millions sont dus à une erreur humaine, et il est probable qu'un trafic fait de voitures autonomes s'avère plus sûr qu'un trafic de voitures conduites par des personnes. Pour l'instant, toutes les voitures autonomes d'un même type sont livrées avec les mêmes paramètres et n'apprennent plus une fois en service, mais on peut augurer qu'à l'avenir elles continueront d'apprendre en continu à partir de leur environnement. L'évaluation périodique de leur comportement deviendra alors cruciale.

II.4 Les robots auprès des personnes et des groupes

L'adaptabilité à l'environnement que confère la capacité d'apprentissage devrait à l'avenir favoriser l'usage des robots auprès des personnes, notamment comme compagnons ou soignants. Construire des robots "sociaux" d'assistance aux personnes nécessite d'encadrer leur usage, d'autant plus quand ils sont en contact avec des personnes malades ou âgées.

18http://www.lemonde.fr/pixels/article/2017/04/15/quand-l-intelligence-artificielle-reproduit-le-sexisme-et-le-racisme-des-humains_5111646_4408996.html
19https://hal.inria.fr/ALLISTENE-CERNA/hal-01086579v1, 2014
20 Par exemple http://www.lexpress.fr/actualite/monde/amerique-nord/premier-accident-mortel-pour-une-voiture-tesla-en-pilote-automatique_1808054.html et http://www.lefigaro.fr/secteur/high-tech/2016/03/01/32001-20160301ARTFIG00118-la-google-car-provoque-son-premier-accident-de-la-route.php

III. Questions éthiques

Les théories éthiques traditionnelles trouvent une illustration nouvelle dans le numérique et l'apprentissage machine. Les dilemmes liés aux voitures autonomes en sont une illustration abondamment commentée[21]. De façon caricaturale, un véhicule autonome qui se trouverait à choisir entre sacrifier son jeune passager, ou deux enfants imprudents, ou un vieux cycliste en règle, pourrait être programmé selon une éthique de la vertu d'Aristote – ici l'abnégation - s'il sacrifie le passager, selon une éthique déontique de respect du code de la route s'il sacrifie les enfants, et selon une éthique conséquentialiste s'il sacrifie le cycliste – ici en minimisant le nombre d'années de vie perdues.
Le propos n'est pas ici de traiter de telles questions qui relèvent de la société toute entière, mais d'étudier à l'attention du chercheur, dans le cas de l'apprentissage machine, certaines propriétés particulières que doit satisfaire le comportement d'un dispositif numérique.
Pour tout système numérique, il faut viser à satisfaire les propriétés présentées en III.1. Les systèmes d'apprentissage machine présentent des spécificités évoquées en III.2 qui entrent en tension avec ces propriétés générales.

III.1 Les propriétés générales des systèmes numériques

- **La loyauté et l'équité :** la loyauté des systèmes informatiques signifie que ces systèmes se comportent, au cours de leur exécution, comme leurs concepteurs le déclarent. Si par exemple, ceux-ci déclarent qu'un système n'archive pas les données personnelles de ses utilisateurs, il ne doit pas le faire. L'équité d'un système informatique consiste en un traitement juste et équitable des usagers.

- **La transparence, la traçabilité et l'explicabilité :** la transparence d'un système signifie que son fonctionnement n'est pas opaque, qu'il est possible, par exemple pour un utilisateur, de vérifier son comportement. Cette transparence s'appuie notamment sur la traçabilité, la mise à disposition d'informations sur ses actions suffisamment détaillées (mémorisées dans un journal) pour qu'il soit possible après coup de suivre ses actions. La traçabilité est essentielle d'une part pour les attributions de responsabilité, où elle permet le cas échéant de fonder un recours juridique, et d'autre part pour diagnostiquer et corriger les dysfonctionnements. Elle permet également d'expliquer le fonctionnement d'un système à partir des traces laissées par celui-ci ; on parle d'explicabilité.

- **La responsabilité** : la possibilité d'identifier des responsabilités lors d'un dysfonctionnement implique la possibilité de distinguer deux agents : un concepteur et un utilisateur du système. Le donneur d'ordre ou le concepteur est responsable si le système est mal conçu, l'utilisateur est responsable s'il a mal utilisé le système (de même que lors de l'utilisation d'un marteau, l'utilisateur est responsable si, par maladresse, il se tape sur les doigts alors que le concepteur est responsable si la cognée se détache du manche et assomme l'utilisateur) tout en sachant que le professionnel est tenu à un devoir d'information supplémentaire vis à vis de tout non professionnel (l'usager).

21 Jean-François Bonnefon, Azim Shariff2, Iyad Rahwan, The social dilemma of autonomous vehicles, Science 24 Jun 2016

- **La conformité** : Un système numérique doit demeurer conforme à son cahier des charges, et son cahier des charges doit être conforme à la législation. La conformité d'un système à son cahier des charges signifie que le système est conçu pour effectuer des tâches spécifiées en respectant les contraintes explicitées dans ce cahier. Les spécifications du cahier des charges formalisent souvent une interprétation restrictive de la loi, à défaut de pouvoir traduire les finesses de cette dernière. La conformité doit être vérifiée avant que le système soit utilisé, en analysant son code et ses données. Par exemple, la conformité inclut le respect des règles de protection des données personnelles pour un système d'analyse de données ou le respect du code de la route pour la voiture autonome.

III.2 Quelques spécificités des systèmes d'apprentissage machine

- **Difficulté de spécification :** Le but de l'apprentissage automatique est précisément de traiter des tâches que l'on ne sait pas spécifier au sens informatique. L'apprentissage automatique est une façon de remplacer cette spécification formelle par un modèle paramétré empiriquement par la machine à partir d'une masse de données. Pour concevoir un programme qui recommande des livres à des lecteurs dans une bibliothèque publique, deux types d'approches sont possibles. Le premier utilise des règles explicites : un tel programme peut par exemple comporter trois listes de livres destinés aux enfants, aux adolescents et aux adultes, demander son âge à une lectrice et selon que cet âge est inférieur à 12 ans, compris entre 13 et 17 ans ou supérieur à 18 ans, tirer au hasard un livre dans l'une de ces listes. Dans ce cas, il est facile de le spécifier : il faut que le livre recommandé soit dans la catégorie qui correspond à l'âge de la lectrice. Le second utilise un algorithme d'apprentissage qui procède différemment : il se fonde sur l'âge de la lectrice et sur la liste des livres que les lecteurs de son âge ont déclaré avoir appréciés. La liste des livres recommandés à chacune des classes d'âges est alors dynamique : elle varie au cours de l'entraînement de l'algorithme. Un avantage est qu'au lieu de s'appuyer sur un découpage grossier de la population en trois grandes catégories, les recommandations peuvent être beaucoup plus fines. Un inconvénient est qu'un entraîneur facétieux peut entraîner l'algorithme afin qu'il recommande les livres inappropriés à l'âge des enfants. Dans ce cas, puisque les livres ne sont pas catégorisés *a priori*, il est impossible de donner un sens à la spécification : « Le livre recommandé doit appartenir à la catégorie qui correspond à l'âge de la lectrice ». De plus, nous supposons dans cet exemple que, si l'algorithme d'apprentissage construit dynamiquement la liste des livres à recommander aux lecteurs en fonction de leur âge, il construit également les « catégories » utilisées pour établir lesdites recommandations. Il peut ainsi utiliser, non les concepts habituels (âge, genre, etc.), mais des concepts qui lui sont propres, dont l'humain ne comprendrait pas nécessairement la signification, ce qui rend plus difficile encore l'expression de la spécification de ce que nous attendons de l'algorithme. Par exemple, la catégorie des « lecteurs qui demandent un prêt de 15h à 15h15 » peut être pertinente pour la machine tout en paraissant aléatoire ou dénuée de sens pour un utilisateur humain. Dans le cas d'une bibliothèque publique, l'explicabilité des recommandations est impérative. Elle nécessite que les catégories conduisant à un résultat, même si elles émergent d'un processus d'apprentissage, soient exprimées dans le langage humain et clairement spécifiées.

- **Agent entraîneur** : Outre le concepteur et l'utilisateur, les systèmes d'apprentissage machine introduisent un troisième type d'agent qui entraîne le système d'apprentissage machine avec un jeu de données. Un dysfonctionnement du système peut désormais être dû à une mauvaise conception ou une mauvaise utilisation du système, auxquels cas le concepteur ou l'utilisateur doivent être tenus pour responsables, mais aussi à un mauvais entraînement du système, auquel cas c'est l'entraîneur qui doit être considéré comme responsable. Cette situation n'est pas complètement nouvelle. Quand l'auteur d'un programme utilise un compilateur, trois agents entrent également en jeu : le concepteur du compilateur, homologue du concepteur du système d'apprentissage, l'auteur du programme, homologue de l'entraîneur et l'utilisateur du programme, homologue de l'utilisateur du système d'apprentissage. Dans cette situation le programme transforme et est transformé : il transforme les entrées, mais il est transformé par le compilateur. Toutefois la réalité est plus complexe encore. Dans un système qui apprend en continu, tous les utilisateurs sont aussi des entraîneurs. En amont de l'entraîneur ou avec son accord, des données peuvent avoir été obtenues à partir de questions sensibles ou soumises à des restrictions ou interdictions de traitements (données personnelles, droit à l'image, etc).

- **Apprendre sans comprendre :** Les algorithmes d'apprentissage automatique permettent de battre les meilleurs joueurs aux échecs ou au Go, mais ils sont souvent incapables d'expliquer la raison pour laquelle ils ont joué un coup et non un autre, car cette « explication » se fonde sur le réglage de millions de poids synaptiques et non sur des concepts simples et intelligibles par les humains[22]. De même, l'une des forces des algorithmes de reconnaissance d'image par apprentissage est de reconnaître une chaise sans nécessairement utiliser les concepts de pied, d'assise, de dossier… de ce fait, l'algorithme peut difficilement expliquer les raisons pour lesquelles il a identifié une chaise dans une image. Le problème est plus complexe encore dans le cas d'un algorithme non supervisé, car l'algorithme apprend sans référence à un but intelligible par les humains, ou dans le cas d'un algorithme par renforcement, qui vise à optimiser une fonction de récompense souvent trop simple pour permettre d'expliquer comment le but affiché est atteint. Les corrélations entre les concepts appris (les *clusters* ou le vocabulaire d'indexation) et les zones d'une image analysée diffèrent parfois fortement entre l'homme et la machine : le réseau et l'humain ne privilégient pas les mêmes zones de l'image pour répondre à une question sur cette image[23].

- **Evolution dynamique du modèle :** lorsque le système continue à apprendre après son déploiement, il est difficile de contrôler son comportement sur le long terme. Le système peut apprendre au cours de son utilisation des comportements induisant un biais. Le système est alors non-équitable. Il peut s'agir, par exemple, d'un comportement abusif pour un robot au contact d'humains ou d'un algorithme faisant des offres financières différentes à des minorités ou groupes sociaux particuliers. De même, l'algorithme lui-même peut faire émerger des catégories inattendues qui peuvent s'avérer discriminatoires sur le fondement des libertés fondamentales (critères non significatifs de sélection de risque en

22 Un joueur humain de go peut également ne pas être capable d'expliquer un coup. Dans le cas des jeux, ce manque d'explicabilité est acceptable.

23 Human Attention in Visual Question Answering : Do Humans and Deep Networks Look at the Same Regions? Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh,Batra, 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, arXiv :1606.03556

matière de crédit *scoring*, comme la taille du demandeur). Le mode de recueil des données utilisées pour l'apprentissage n'est pas toujours facile à repérer.

- **Instabilité de l'apprentissage :** l'apprentissage profond est une des approches les plus performantes à l'heure actuelle, pourtant cet algorithme peut montrer une certaine instabilité. La modification imperceptible à l'œil d'un petit nombre de pixels dans une image identifiable par un humain peut rendre cette image non identifiable par un système d'apprentissage profond, par exemple la photo d'une voiture[24]. A l'inverse, certaines images qui n'ont aucune signification pour les humains sont parfois étiquetées automatiquement comme proches de formes apprises. Des images ayant pour les humains des sens très différents peuvent être indexées de manière identique : un exemple classique est une photo de panda reconnue comme gibbon[25]. Les valeurs de sortie dans un réseau profond associent un degré de confiance à une reconnaissance ; une méthode de gradient permet alors d'augmenter cette confiance en gardant les paramètres du réseau mais en modifiant pas à pas l'entrée, convergeant ainsi vers une entrée donnant la sortie considérée avec un maximum de confiance. La plate-forme *Deep Dream* de Google permet facilement de transformer ainsi des photos, dont on constate qu'elles prennent des allures hallucinatoires.

- **Evaluation et contrôle :** Comme il est difficile de formuler le cahier des charges d'un système qui apprend, il est difficile d'évaluer ce système. Il est en revanche possible d'en évaluer les effets *a posteriori*. Par exemple, il est difficile d'évaluer si un véhicule autonome accélère ou freine au bon moment, mais il est possible d'évaluer *a posteriori* si ce véhicule a provoqué moins ou plus d'accidents qu'un véhicule conduit par un être humain. Lorsque les systèmes qui apprennent continuent à évoluer, leur évaluation doit se répéter à intervalles réguliers tout au long de la période pendant laquelle ces systèmes sont utilisés. Différents types d'agents pourraient apparaître dans la gestion des systèmes apprenants : des agents « interpréteurs » qui aident à comprendre le comportement de la machine à partir de jeux de test, des agents « évaluateurs » ou « vérificateurs » des algorithmes d'apprentissage afin de vérifier que les systèmes apprenants restent loyaux et équitables, des agents « juges » qui vérifient que ces systèmes se comportent conformément à la loi lors de leur utilisation.

Ces spécificités, qui sont autant de points de vigilance pour le chercheur, sont la contrepartie de l'ampleur des possibilités offertes par l'apprentissage machine. Elles ouvrent le champ à de nouvelles recherches qui permettront dans de nombreux secteurs de développer des systèmes destinés à être plus sûrs que ne le sont les humains, êtres apprenants qui eux aussi faillissent et se trompent.

IV. Les préconisations sur les systèmes apprenants en six thèmes

IV.1 Les données des systèmes d'apprentissage

Les données conditionnent le résultat de l'algorithme d'apprentissage. Leur captation par apprentissage en milieu ouvert (c'est-à-dire dans des conditions réelles d'utilisation) et l'absence de modèle a priori rendent difficile l'appréciation de leur adéquation aux

24 Intriguing properties of neural networks, Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus https://arxiv.org/abs/1312.6199

25 Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, Explaining and harnessing adversarial examples, ICLR 2015 https://arxiv.org/pdf/1412.6572v3.pdf

objectifs, et les biais des données, intentionnels ou non, peuvent être lourds de conséquences **[DON-1]**. Les données d'apprentissage peuvent être discriminatoires, favorisant certaines caractéristiques, comme dans le cas de la reconnaissance faciale, pouvant refléter les préférences politiques, religieuses, ou autres de ses concepteurs ou entraîneurs. Par exemple, Google avait mis un système en ligne en 2015 qui permettait la reconnaissance des visages et fonctionnait mieux sur les peaux claires. Google avait dû s'en excuser publiquement[26]. Dans certains cas, les biais peuvent être illégaux, comme, par exemple, d'offrir des produits financiers moins avantageux pour les membres de groupes minoritaires. A titre d'exemple de partialité commerciale, l'Union Européenne a condamné récemment Google, accusée de biaiser les résultats de son moteur de recherche en faveur de ses propres produits[27].

L'équité est difficile à spécifier. Elle nécessite une appréciation humaine. Elle peut mettre en œuvre des données sur les personnes selon des critères empreints de subjectivité ou conflictuels. Selon les cultures, elle peut par exemple viser prioritairement l'égalité ou la récompense des mérites (la notion « d'égalité des chances » illustre la complexité). Sur le plan individuel, si l'équité consiste à servir chacun au mieux de ses aspirations, un robot doit-il faire plaisir à une personne âgée en lui servant un whisky, puis un deuxième, même si cela n'est pas bon pour sa santé ? **[DON-2]**.

La loi peut interdire l'utilisation de certaines variables comme l'ethnie, le sexe, l'âge, la religion pour classer des individus, afin par exemple de leur accorder ou pas certains services personnalisés. Pourtant, un algorithme peut parfois reconstruire les valeurs de telles variables, et ensuite prendre des décisions fondées sur ces valeurs **[DON-3]**.

Le traçage assure la disponibilité de « suivis » des données captées dans l'environnement ou échangées lors de certains événements, ainsi que des calculs réalisés par le système. Il est indispensable à la transparence et à l'analyse du fonctionnement ou des dysfonctionnements. Si le code du système et ses données sont ouverts - deux éléments clés de la transparence - le contrôle est évidemment plus facile (même s'il reste difficile à réaliser). Dans le cas de machines apprenantes, pour la phase précédant la mise en service il convient autant que possible de conserver les données d'apprentissage, les conditions de leur collecte et de leur validation. Si le dispositif continue d'apprendre en cours d'exploitation, le traçage est plus compliqué. Il est indispensable de monitorer les traces pour éventuellement détecter des déviations par rapport aux comportements attendus. Dans de tels cas, le système peut demander à un humain de valider l'état de l'apprentissage. Le chercheur aura à l'esprit que les traces d'apprentissage sont des données et qu'à ce titre elles doivent respecter les règles concernant les données à caractère personnel, même si elles sont destinées aux seules fins de contrôle technique **[DON-4]**.

Points d'attention et préconisations

[DON-1] Qualité des données d'apprentissage

Le concepteur et l'entraîneur veilleront à la qualité des données d'apprentissage et des conditions de leur captation tout au long du fonctionnement du système. Les entraîneurs du système informatique sont responsables de la présence ou de l'absence de biais dans les données utilisées dans l'apprentissage, en particulier l'apprentissage « en continu », c'est-

26 http://www.huffingtonpost.fr/2015/07/02/logciel-reconnaissance-faciale-google-confond-afro-americains-gorilles_n_7711592.html

27 http://europa.eu/rapid/press-release_IP-16-2532_fr.htm

à-dire en cours d'utilisation du système. Pour vérifier l'absence de biais, ils doivent s'appuyer sur des outils de mesure qui restent encore à développer.

[DON-2] Les données comme miroir de la diversité

Les entraîneurs des systèmes d'apprentissage automatique doivent opérer le choix des données en veillant à ce que celles-ci respectent la diversité des cultures ou des groupes d'utilisateurs de ces systèmes.

[DON-3] Variables dont les données comportent un risque de discrimination

Les entraîneurs (qui peuvent être aussi les concepteurs ou les utilisateurs) doivent se poser la question des variables qui peuvent être socialement discriminantes. Il convient de ne pas mémoriser ni de régénérer par programmation ces variables, par exemple l'ethnie, le sexe ou l'âge. La protection des données à caractère personnel doit également être respectée conformément à la législation en vigueur.

[DON-4] Traces

Le chercheur doit veiller à la traçabilité de l'apprentissage machine et prévoir des protocoles à cet effet. Les traces sont elles-mêmes des données qui doivent à ce titre faire l'objet d'une attention sur le plan éthique.

IV.2 L'autonomie des systèmes apprenants

Pour un système numérique, « l'autonomie est sa capacité à fonctionner indépendamment d'un opérateur humain ou d'une autre machine en exhibant des comportements non triviaux dans des environnements complexes et variables.(...) [28]» L'autonomie est un concept relatif. L'autonomie que peut atteindre un système dépend d'une part de la complexité de l'environnement, qui peut être mesurée par la quantité et la variabilité d'information et de son flux et de sa dynamique, et d'autre part de la complexité de la tâche, qui dépend de la structure de l'ensemble des états possibles du système (espace d'états). Si l'environnement d'utilisation d'un système autonome est complexe, comme les scènes de la rue pour une voiture autonome, un apprentissage préalable est souvent nécessaire. Si les environnements d'utilisation sont évolutifs ou imprévisibles, comme pour un robot compagnon, un apprentissage personnalisé doit être réalisé, qu'il peut être nécessaire d'actualiser périodiquement ou de continuer tout au long de l'exploitation.

Le fonctionnement fidèle (à ce que les concepteurs et les opérateurs ou utilisateurs attendent de lui) d'une machine dotée d'autonomie nécessite que la représentation d'une situation et le comportement calculés par la machine soient intelligibles et conformes à ce que son opérateur ou utilisateur humain en attend. Des préconisations sont formulées à cet égard dans l'avis *Éthique de la recherche en robotique* de la CERNA. Si la machine est dotée de capacités d'apprentissage, l'instabilité de l'apprentissage, les corrélations inattendues qu'il peut induire, font que les représentations internes à la machine de la situation et ses plan d'action peuvent être sans rapport avec ce que l'utilisateur imagine **[AUT-1]**.

D'une façon générale, l'apprentissage peut permettre d'étendre l'autonomie de la machine quant à la façon d'atteindre le but qui lui est assigné. Ainsi *AlphaGo* s'est amélioré en jouant contre des copies de lui-même : cet apprentissage par renforcement illustre une possibilité d'évolution par sélection, sans intervention humaine, des

systèmes apprenants où seuls les plus compétitifs sont dupliqués pour de nouveaux affrontements. Selon Nick Bostrom[29], la machine pourrait évaluer qu'il est plus efficace de s'affranchir de contrôles humains, et pour cela apprendre à dissimuler sa stratégie et neutraliser les reprises en main, elle pourrait même faire émerger des objectifs se substituant à la finalité qui a guidé sa conception. Une telle hypothèse ne repose actuellement sur aucune base scientifique mais elle nourrit la science-fiction et les effets médiatiques, qui brouillent trop souvent la frontière entre la réalité scientifique et les fantasmes. Le scientifique doit tenir compte de cette situation dans sa communication **[AUT-2].**

Points d'attention et préconisations

[AUT-1] Biais de caractérisation
Le chercheur veillera à ce que les capacités d'apprentissage d'un système informatique n'amènent pas l'utilisateur à croire que le système est dans un certain état lors de son fonctionnement alors qu'il est dans un autre état.

[AUT-2] Vigilance dans la communication
Dans sa communication sur l'autonomie des systèmes apprenants relativement aux humains, le chercheur doit viser à expliquer le comportement du système sans donner prise à des interprétations ou des médiatisations irrationnelles.

IV.3 L'explicabilité des méthodes d'apprentissage et leur évaluation

L'exigence d'explications, codifiée à travers la gestion du risque dans les secteurs traditionnels de l'industrie et par les règles de certaines professions (médecine, droit), s'affirme dans le numérique, où certains aspects s'inscrivent dans les textes législatifs (loi Informatique et Libertés, loi sur la République numérique).

Expliquer un algorithme est faire comprendre à ses utilisateurs ce qu'il fait, avec assez de détails et d'arguments pour emporter leur confiance. Cette tâche est difficile même dans le cas d'un algorithme dépourvu de capacité d'apprentissage, comme l'illustre le débat autour de l'algorithme d'admission post-bac APB[30]. En outre il convient de distinguer preuve et explication : ainsi Gilles Dowek donne l'exemple simple de la multiplication de 12345679 par 36, dont le seul calcul du résultat (444444444) n'explique pas aux yeux d'un esprit mathématique pourquoi ce résultat ne comporte que des 4.

Pour qu'un algorithme soit explicable, ses principes doivent être suffisamment documentés pour être compréhensibles par l'ensemble des usagers, avec la médiation éventuelle d'experts ; le passage de l'algorithme au code puis l'exécution du programme doivent être formellement vérifiés, ce qui est affaire de spécialistes. En fin de comptes, l'explicabilité d'un algorithme repose sur des méthodes rigoureuses mais aussi sur un corpus de connaissances non formalisées partagées entre les humains.

La capacité d'apprentissage accroît considérablement la difficulté d'explication, et fait que le concepteur lui-même peut ne pas être en mesure de comprendre le comportement du système [31]. En effet, alors que les algorithmes classiques traduisent

29 Superintelligence, Oxford University Press, 2014

30 Rapport de la mission Etalab sur les conditions d'ouverture du système Admission Post-Bac , avril 2017

31 The Dark Secret at the Heart of AI, Will Knight, MIT Technology Review, avril 2017 https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

un modèle qui se prête aux explications parce que fourni par des analystes, l'apprentissage calcule un modèle interne par ajustement de ses paramètres, parfois des millions, en fonction de données qui pour nous ont un sens mais qui ne sont pour la machine que des suites d'octets, au risque d'une grande instabilité d'interprétation et de corrélations inattendues. Les difficultés à s'assurer du comportement d'une machine apprenante, et *a fortiori* à l'expliquer, illustrées en III.2 dans le cas de l'apprentissage supervisé, grandissent dans le cas de l'apprentissage par renforcement ou la classification, où les données d'apprentissage ne sont plus annotées par l'humain.

Un compromis entre les capacités d'apprentissage et l'explicabilité s'avère nécessaire. Ce compromis est à apprécier en fonction des domaines d'applications : si l'explicabilité n'est *a priori* pas une exigence dans des applications telles le jeu, elle est impérative dès lors que l'on touche aux intérêts et droits des personnes ou à leur sécurité. Le chercheur devra assurer et documenter un niveau d'explicabilité acceptable selon le type d'application, et notamment en expliciter les limitations et les médiations d'experts requises **[EXP-1]**.

Des méthodes nouvelles d'explication du fonctionnement et des résultats des machines apprenantes voient le jour, visant à améliorer ce compromis. La DARPA en a même fait en 2016 un appel à projet spécifique[32]. Ces méthodes peuvent comporter des heuristiques ou des outils d'observation, comme la visualisation du comportement, qui n'ont pas valeur d'explication conceptuelle, et le chercheur veillera alors à ne pas en tirer hâtivement des catégorisations des données, au risque de laisser place à des biais, y compris des biais idéologiques ou politiques **[EXP-2]** - par exemple, l'interprétation en termes anthropométriques de l'observation d'un système de reconnaissance faciale**.**

Le besoin d'évaluation (conformité, équité, loyauté, neutralité, transparence...) des plate-formes et des algorithmes devient un sujet de société, objet de débats et de réglementations (voir V). Des normes, des procédures d'évaluation pour la mise sur le marché et les contrôles en exploitation vont ainsi émerger pour concourir à une bonne gouvernance des algorithmes. Le chercheur doit prendre en compte cette évolution et contribuer au débat public et à l'élaboration de normes et de pratiques d'évaluation et de recours **[EXP-3].**

L'évaluation des machines apprenantes est un sujet scientifique largement ouvert. Avec un programme classique, la vérification du code est déjà un problème extrêmement compliqué. Pour un système apprenant, la tâche est mouvante puisque des erreurs, des biais, des comportements inacceptables peuvent surgir dans le temps, comme cela a été illustré dans le cas de l'apprentissage profond. Une grande difficulté est de trouver des critères mesurables, et des échantillons tests de situations garantissant un fonctionnement correct dans l'ensemble des situations d'exploitation. Quand le dispositif apprend en continu en cours d'exploitation, la difficulté est accrue par le fait qu'en milieu ouvert, des situations imprévues, aux conséquences elles-mêmes imprévisibles peuvent être rencontrées.

Comme pour d'autres domaines, il est nécessaire d'explorer des procédures d'autorisation de mise sur le marché et de contrôle « technique ». L'idée a été émise de tests réalisés périodiquement par une agence indépendante de contrôle. Cette piste semble difficile à mettre en œuvre. D'une part, les difficultés techniques évoquées ci-dessus devraient être surmontées. D'autre part, tous les dispositifs concurrents devront

32 http://www.darpa.mil/program/explainable-artificial-intelligence

être testés en même temps sur une même batterie de tests non dévoilés à l'avance, ce qui est difficilement réalisable. Enfin le risque est gros que les machines soient conçues pour satisfaire aux tests. L'actualité de début 2016 - dans un contexte beaucoup plus simple concernant le paramétrage par les constructeurs automobiles des moteurs en vue des tests de pollution - alimente cette réserve. Le plan stratégique en recherche et développement en intelligence artificielle de la Maison Blanche[33] préconise pour sa part une panoplie de mesures basées sur des infrastructures ouvertes de développement, de test et d'évaluation, comprenant la collecte et l'ouverture de données publiques massives et d'environnements logiciels.

Points d'attention et préconisations

[EXP-1] Explicabilité
Le chercheur doit s'interroger sur la non-interprétabilité ou le manque d'explicabilité des actions d'un système informatique apprenant. Le compromis entre performance et explicabilité doit être apprécié en fonction de l'usage et doit être explicité dans la documentation à l'usage de l'entraîneur et de l'utilisateur.

[EXP-2] Les heuristiques d'explication
Dans sa recherche d'une meilleure explicabilité d'un système apprenant, le chercheur veillera à décrire les limitations de ses heuristiques d'explication et à ce que les interprétations de leurs résultats ces soient exemptes de biais.

[EXP-3] Elaboration des normes
Le chercheur veillera à contribuer aux débats de société et à l'élaboration de normes et de protocoles d'évaluation qui accompagnent le déploiement de l'apprentissage machine. Quand les données concernent certains secteurs professionnels spécialisés (médecine, droit, transports, énergie, etc.), les chercheurs de ces domaines devront être sollicités.

IV. 4 Les décisions des systèmes d'apprentissage

Depuis le système expert médical Mycin des années 1970, des dispositifs d'aide à la décision existent dans de nombreux domaines, dont le secteur régalien du droit et celui vital de la santé. La question de la place à accorder dans le processus de décision aux propositions fournies par la machine se pose de plus en plus. Les décisions graves, comme condamner une personne à la prison, sont encore prises par des êtres humains. Mais une multitude de décisions aux conséquences moindres (condamner un automobiliste à une amende, accorder ou refuser un prêt à la consommation, etc.) sont déjà prises par des algorithmes. Une personne humaine responsable demeure encore associée à toutes les décisions, ouvrant la possibilité de faire un recours, mais la tendance est clairement à l'automatisation.

Les décisions des machines peuvent s'avérer plus sûres et moins biaisées que celles des humains et de leurs humeurs. Dans certains cas la rapidité de décision des machines peut même s'avérer déterminante. La démarche éthique n'est pas de se priver de tels avantages, mais de prendre conscience que, d'une part l'apprentissage est de nature à réintroduire de l'insécurité et des biais, et d'autre part ces avantages sont à articuler avec la perception qu'en a l'humain - un patient pourrait mieux accepter une erreur de son médecin que d'une machine. La difficulté pour l'humain de contester la décision de la machine en ne restituant pas la part discrétionnaire présente dans la plupart des

33https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf , octobre 2016.

décisions humaines est à considérer (cf article 22 du règlement européen sur la protection des données[34]).

- La personne qui décide risque de n'être plus qu'un exécutant de la « proposition » formulée par la machine. Suivre cette proposition, s'aligner sur la décision de la machine apparaît comme l'option la plus sûre. Dévier de la solution proposée par la machine est un acte qui nécessite d'être expliqué, qui entraîne une prise de responsabilité et de risque.
- La personne dont le sort dépend d'une décision automatisée risque pour sa part d'être réduite à un profil et de ne pas pouvoir rendre compte de sa situation individuelle, de ses motifs, de ses raisons, bref de sa singularité.

Dans les deux cas se pose la question de la capacité des personnes à accomplir des actes et à rendre compte de ceux-ci.

La capacité d'apprentissage des programmes risque d'amplifier cette tendance en rendant les propositions de décisions circonstanciées et individualisées. Le manque d'explicabilité des résultats et leur variabilité **[DEC-1]** (selon les données apprises antérieurement et leur chronologie) ne doivent nullement pouvoir être assimilés à un « pouvoir discrétionnaire de la machine » mais imposent au contraire

- d'affirmer la primauté de la décision et de l'explication humaines, par exemple par une obligation de justifier la décision au regard de la personne qu'on a devant soi ;
- de travailler à la transparence et la loyauté des algorithmes utilisés, et à encadrer par la loi leur validation et leur évaluation.

Comme dans d'autres secteurs, l'introduction de l'apprentissage machine dans l'aide à la décision induit une montée en qualification des métiers concernés, voire l'émergence de nouveaux métiers et la disparition d'autres. Les tâches fastidieuses de recherche documentaire et de jurisprudence sont déjà déléguées à des machines dans des cabinets d'avocats. En contrepartie le sujet humain doit être formé à la compréhension et l'interprétation des résultats de la machine, et se consacrer à communiquer et expliquer le sens des décisions. Le concepteur des machines apprenantes d'aide à la décision doit être impliqué dans l'évolution de l'environnement réglementaire et humain qui résulte de leur utilisation **[DEC-2]**.

Points d'attention et préconisations

[DEC-1] Place de l'humain dans les décisions assistées par des machines apprenantes
Le chercheur, concepteur de machines apprenantes d'aide à la décision, doit veiller à ce qu'aucun biais ne produise un résultat qui devienne automatiquement une décision alors que l'intervention humaine était prévue par la spécification du système, et prendra garde aux risques de dépendance de l'humain aux décisions de la machine.

[DEC-2] Place de l'humain dans l'explication des décisions assistées par des machines apprenantes
Le chercheur doit veiller à ce que les résultats du système soient autant que possible interprétables et explicables pour les personnes concernées, s'impliquer dans les nécessaires évolutions des métiers en faisant usage et dans l'encadrement des pratiques.

34 https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre3#Article22

Des agents experts contribueront à l'explication et à la vérification des comportements du système.

IV.5 Le consentement à l'apprentissage machine

L'utilisation de systèmes apprenants interconnectés fait émerger une exigence de consentement portant sur les conséquences que les capacités d'apprentissage de ces systèmes peuvent entraîner pour les individus et pour les groupes.

Nous consentons aujourd'hui à ce que des données sur notre comportement soient captées par des objets connectés (de l'ordinateur au robot) afin de nous rendre des services. Ces services dépendent parfois de paramètres évolutifs calculés par apprentissage sur de grandes quantités de données de nature diverse, selon des finalités qui peuvent ne pas être explicites. Il est possible que le concepteur lui-même mésestime l'impact de son application sur l'environnement numérique global. L'impossibilité d'informer l'utilisateur de façon sûre ou précise sur les finalités est due à la donne technique et algorithmique, et plus concrètement, au fait que l'apprentissage peut résulter en des configurations du système que le concepteur ne pouvaient pressentir. Cette situation est nouvelle par rapport au cas du consentement pour un usage ou un type d'usage spécifique.

A titre d'exemple, l'utilisation d'un agent conversationnel qui s'adapte par apprentissage aux habitudes des utilisateurs illustre comment un tel système peut rétroagir de façon implicite sur le comportement de ces utilisateurs. Il peut par exemple imiter la parole de l'utilisateur jusqu'à répéter ses tics verbaux, ce qui risque de perturber l'interaction. Il est nécessaire que les usagers aient à expliciter un consentent à utiliser des machines qui ont la possibilité d'adaptation et soient vigilants sur des comportements non souhaitables de la machine. Pour des personnes fragiles (personnes âgées, enfants) notamment, il convient d'éviter la perturbation par une machine dont le comportements évolue significativement sans que l'utilisateur soit mis au courant de cette possibilité. Les usagers devraient avoir la possibilité de se servir ou non de la fonction d'apprentissage et de contrôler au moins globalement les données que la machine utilise en apprenant : les leurs, celles collectées sur le réseau ou toutes données disponibles **[CON-1]**.

Au niveau des groupes, des sociologues et philosophes étudient l'impact au travail d'un environnement d'intelligence artificielle, notamment en matière de mérite et d'évaluation des performances[35]. Derrière les vertus facilitatrices de certains dispositifs, peut se déployer de manière sous-jacente une normativité prenant la forme d'un paternalisme technologique aux multiples visages: ainsi, les environnements d'intelligence artificielle peuvent alerter, recommander, rappeler à l'ordre, bloquer, interdire, voire influencer. Il s'agit alors de penser les effets potentiels de la technologie sur les capacités et l'autonomie des individus, notamment en considérant les possibilités d'improvisation et de spontanéité[36].

De même le droit à l'oubli ou au retrait numérique — notamment le droit d'une personne de demander au moment où elle retire son consentement, à ce que toutes les

35N. Daniels, « Merit and Meritocracy », *Philosophy and Public Affairs*, Vol. 7, No. 3, 1978, pp. 207-208 : « Merit is construed as ability plus effort ».

360. McLeod, « Desert », in *Stanford Encyclopedia of Philosophy*, *First published Tue May 14, 2002*, *substantive revision Wed Nov 12, 2008, p. 2,* http://plato.stanford.edu/entries/desert/.

données précédemment prélevées sur elle soient effacées – peut devenir illusoires dans la mesure où ses données ont alimenté par apprentissage l'évolution de paramètres censés capter des comportements collectifs.

Plus généralement, il s'agit ainsi d'informer l'usager pour qu'il participe de façon éclairée à la transformation de la société, tout en sachant que dans de telles situations complexes, le consentement ne se fonde pas seulement sur la compréhension rationnelle, mais aussi sur la confiance et, pour une application informatique, la curiosité de l'utilisateur, qui peut être attisée par une volonté de séduction de la part du concepteur**.**

Le chercheur doit délibérer dès la conception de son projet avec les personnes ou les groupes identifiés comme pouvant être influencés, afin que le projet au stade futur d'utilisation puisse recueillir le consentement des parties concernées **[CON-2].** Cette préconisation rejoint une recommandation générale de la CERNA sur la conduite de projet **[GEN-7]**.

Il s'agit plus largement d'être conscient que l'apprentissage machine contribue à déplacer le consentement du niveau individuel de l'usage de ses données personnelles, à un niveau collectif du consentement à ce que des systèmes informatiques puissent servir à orienter la société à partir d'observations globales de cette société. Des recherches dans ce domaine pourraient conduire à des dispositifs nouveaux concernant l'apprentissage machine **[CON-3].**

Points d'attention et préconisations

[CON-1] Choix des usagers d'activer ou non les capacités d'apprentissage d'un système
Le chercheur doit créer la possibilité d'utilisation de systèmes avec ou sans leur capacité d'apprentissage. Il doit donner à l'utilisateur au moins un paramètre de contrôle global sur la source des données utilisées pour l'apprentissage.

[CON-2] Consentement dans le cadre de projet
Le chercheur doit délibérer dès la conception de son projet avec les personnes ou les groupes identifiés comme pouvant être influencés.

[CON-3] Consentement à utiliser une machine capable d'apprentissage en continu

Le chercheur doit être conscient que la capacité d'apprentissage et la mise en réseau de telles capacités peut induire des problématiques nouvelles concernant autant le consentement de l'usager que celui de la société.

IV.6 La responsabilité dans les relations homme-machine apprenantes

Le chapitre IV.4 abordait la délégation de décision à la machine du point de vue de son impact sur l'humain. C'est l'aspect juridique qui est ici abordé. Dans le droit actuel, une machine, qu'elle calcule ou pas, est une chose. La responsabilité, elle, revient toujours à une personne. Cette personne responsable peut être le concepteur de la machine, son entraîneur ou son utilisateur. La responsabilité pour risque ou la responsabilité assurantielle s'applique également au producteur ou au vendeur du système informatique en tant qu'objet commercial.

La première question consiste à distinguer à laquelle ou lesquelles de ces trois catégories d'agents attribuer une responsabilité dans le cas des systèmes informatiques capables d'apprendre. Des lignes directrices sont nécessaires, permettant d'établir une séparation entre les domaines de responsabilité du concepteur, de l'entraîneur et de l'utilisateur, voire une définition juridique rigoureuse de ces domaines. Ces lignes directrices doivent s'appuyer sur la possibilité de reconstruire la chaîne des décisions prises algorithmiquement, ce qui implique la traçabilité du système. Les avancées technologiques actuelles montrent l'urgence d'adapter notre législation à cette nouvelle réalité.

La connaissance que possède le concepteur en tant qu'auteur du code qui fait fonctionner le système informatique, lui confère un pouvoir et une responsabilité. Cette connaissance est cependant limitée : un système informatique apprend grâce aux données fournies par l'entraîneur ou celles qu'il collecte sans supervision en cours d'utilisation. Le comportement d'un système apprenant peut même, dans certains cas, devenir totalement imprévisible pour le concepteur : c'est la raison pour laquelle, pour toute fin pratique, le pouvoir de ce dernier est borné dès la mise en exécution du code, moment auquel il perd le contrôle du système même s'il en garde la « paternité ». La responsabilité du concepteur doit être limitée en conséquence. Une telle limitation, qui implique le partage de la responsabilité, s'étend également à l'utilisateur qui possède un système informatique (par exemple, un smartphone ou un robot) en tant qu'objet matériel, mais qui, par manque de connaissance de son fonctionnement interne, n'a aucun pouvoir effectif sur ce système malgré le fait d'être son propriétaire. La responsabilité de l'entraîneur s'étend aux données qu'il fournit. Si ces données contiennent des biais, la responsabilité de l'entraîneur peut être engagée. Toutefois, l'entraîneur peut plaider qu'en n'étant pas concepteur, il ne possède pas la connaissance des algorithmes du traitement des données employées par le système. Pour faciliter le bon usage du système apprenant et aider à départager les responsabilités, le concepteur doit prévoir des mécanismes de contrôle **[RES-1]**, de documenter le système et en décrire les limites d'utilisation, y compris les caractéristiques des données que le système doit recevoir pour apprentissage **[RES-2]**.

La seconde question consiste à se demander si le système informatique pourrait se voir lui-même attribuer une responsabilité. Actuellement, la responsabilité des personnes repose sur l'imputabilité de l'acte dont ces personnes seraient responsables ; or l'intelligence artificielle permet à la machine d'atteindre un degré avancé d'autonomie tel que sa décision ne peut être directement imputée à une personne humaine. Le choix est alors entre deux options : soit l'attribution d'une responsabilité à l'humain malgré le manque d'imputabilité (on pourra avoir recours à la responsabilité du fait des choses, ou à la responsabilité pour produits défectueux), soit la création d'un statut juridique intermédiaire pour le système informatique capable de porter une responsabilité. Nous laissons de côté la deuxième possibilité peu réaliste politiquement, mais intéressante tant juridiquement que philosophiquement et, depuis peu, discutée dans certains cercles européens[37]. Toutefois, la société doit dès à présent prendre conscience du fait que, par exemple, les voitures autonomes devront sans doute être dotées d'un statut particulier, dont les détails transparaîtront progressivement à l'expérience, à l'instar du droit des différentes personnes morales qui s'est forgé avec le temps.

37 European Parliament. Directorate-General for Internal Policies Policy. Department C : Citizens' Rights and Constitutional Affairs. Legal Affairs. European Civil Law Rules in Robotics. Résolution du 12 janvier 2017.

La difficulté d'imputabilité d'une action décidée par un système informatique conduit à distinguer plusieurs figures de responsabilité de l'agent humain, limitée ou partagée :

1) Sur le plan de l'intention : un humain (concepteur, entraîneur, utilisateur) a-t-il formé l'intention de voir la machine produire un certain résultat, même si une ou plusieurs caractéristiques de ce résultat n'étaient pas intentionnelles ?
2) Sur le plan de l'action : un humain (concepteur, entraîneur, utilisateur) a-t-il opéré des choix volontaires ou involontaires, par exemple, une sélection des données utilisées lors de l'apprentissage par la machine ?
3) Sur le plan de la prévisibilité et du hasard : un agent (concepteur, entraîneur, utilisateur) a-t-il pu prévoir l'action de la machine dans des conditions raisonnables de l'exécution de ces fonctions ? Quelle est la place de l'aléatoire dans la décision prise par le système ?
4) Les données (par exemple, celles qui sont fournies par l'entraîneur) peuvent être non conformes aux annonces faites, ou non transparentes, incomplètes, non à jour, inexactes voire falsifiées (ou « hackées ») par un tiers, entraînant dans ce dernier cas l'application de la loi sur la fraude informatique et l'intrusion dans les systèmes informatiques (loi du 5 janvier 1988). Il est aussi possible que des raisonnements conduisent à des catégorisations et, partant, à des discriminations illégales à partir de données sensibles, ou même neutres[38].
5) Sur les aspects logiciels, les distinctions conceptuelles « Abstraction-Filtration-Comparaison » qui existent dans la jurisprudence américaine sur les droits d'auteur peuvent s'avérer utiles dans l'analyse du statut social et juridique des algorithmes d'apprentissage[39]. L'étape de l'abstraction vise à séparer l'idée générale, qui ne peut appartenir à personne, de son expression précise, elle protégée par le droit. Pour cela, le code est analysé selon ses niveaux fonctionnels ; chaque niveau est classé soit comme « idée » soit comme « expression ». À l'étape du filtrage, sont exclus : les éléments incontournables imposés par des raisons d'efficacité, car leur caractère privé serait susceptible de créer un monopole d'accès ; les éléments provenant des facteurs extérieurs, comme par exemple les standards ou les règles d'expression ; les éléments dont la provenance relève du domaine public. L'étape de la comparaison consiste à comparer le résidu avec l'œuvre initiale ouvrant la voie à l'attribution de la propriété et de la responsabilité pour le logiciel et les décisions qu'il a prises.

Points d'attention et préconisations

[RES-1] Mécanismes de contrôle
Le chercheur doit adapter et inclure dans le système, des mécanismes de contrôle, automatiques ou supervisés (par la machine ou l'être humain), sur les données, sur le fonctionnement au niveau informatique et sur le raisonnement suivi, afin de faciliter l'attribution de responsabilités au regard du bon fonctionnement ou du dysfonctionnement du système.

[RES-2] Déclaration des intentions d'usage
En documentant un système informatique apprenant, le chercheur doit décrire de manière sincère, loyale et complète, les limites qui lui sont connues de l'imputabilité d'une décision ou d'une action du système soit au code-source soit au processus d'apprentissage. Cette documentation servira de déclaration d'intention de la part du concepteur quant aux

38Par exemple, dans les logiciels de scoring utilisant les BigData, il existe une discrimination objective « calculée » entre personnes de grande taille ou de petite taille.
39http://jolt.law.harvard.edu/digest/copyright/artificial-intelligence-and-authorship-rights

usages du système informatique qu'il envisage comme normaux. L'absence de telle déclaration ou son caractère tardif peuvent engager la responsabilité supplémentaire du concepteur.

V. Contexte national et international

L'apprentissage machine est un des facteurs des progrès actuels des technologies *Big data*, de l'intelligence artificielle et de la robotique. Le fort impact sociétal de ces aspects du numérique est doublé d'une méconnaissance de ses ressorts scientifiques et technologiques. Suite à ce constat, les initiatives de recherche ou de développement dans le numérique de ces derniers temps comportent toujours un volet sur l'éthique, quand elles n'y sont pas entièrement consacrées.

Au niveau international

La mobilisation de la communauté scientifique internationale est illustrée par l'émergence de nouveaux workshops structurants comme « Fairnebop on Data and Algorithmic Transparency » (DAT'16)[40], « Interpretable Machine Learning for Complex Systems »[41] ou « Machine Learning and the Law »[42] à NIPS 2016.

La plus importante association professionnelle internationale du numérique, l'*Institute of Electrical and Electronics Engineers*, a initié l'*IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems* qui a produit fin 2016 un rapport d'étape *Ethically Aligned Design*[43]. Une initiative pilotée par AT&T et Inria regroupe une communauté d'académiques, industriels, décideurs et régulateurs sur la transparence des données personnelles en ligne à travers des activités de recherche[44].

Le plan stratégique en recherche et développement en intelligence artificielle de la Maison Blanche[45] préconise une panoplie de mesures fondées sur des infrastructures ouvertes de développement, de test et d'évaluation, comprenant la collecte et l'ouverture de données publiques massives et d'environnements logiciels. La DARPA[46] a lancé une initiative de recherche intitulée « *explainable artificial intelligence* » (XAI)[47]. Il faut noter également les efforts menés par l'OTRI (Office of Technology Research and Investigation) créé au sein de la FTC (Federal Trade Commission). Cette dernière a publié en janvier 2016 un rapport intitulé *Big Data, a Tool for Inclusion or Exclusion? Understanding the issues*[48]. L'université de Stanford a lancé en 2014 l'initiative *One Hundred Year Study on Artificial Intelligence (AI100)*[49] dont le rapport 2016[50] a été publié en septembre. Il s'agit d'un programme à long terme pour étudier les impacts de l'intelligence artificielle sur les individus et la société avec des préoccupations sur la démocratie, les libertés, l'éthique en plus des considérations technologiques et scientifiques. Ils sont rejoints par

40 http://datworkshop.org/
41 http://www.mlandthelaw.org/
42 https://sites.google.com/site/nips2016interpretml/
43 http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html
44 http://www.datatransparencylab.org/
45 https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf , octobre 2016.
46 Defense Advanced Research Projects Agency
47 http://www.darpa.mil/program/explainable-artificial-intelligence
48 https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf
49 https://ai100.stanford.edu/
50 https://ai100.stanford.edu/sites/default/files/ai_100_report_0901fnlc_single.pdf

plusieurs grands industriels américains qui tentent de construire un « standard » éthique autour des technologies de l'intelligence artificielle[51]. De nombreux instituts de recherche interdisciplinaires ont été créés récemment, principalement dans les pays anglo-saxons, pour réfléchir aux enjeux de l'intelligence artificielle, par exemple l'Institut du futur de l'humain (FHI pour *Future of Human Institute*) d'Oxford et le centre de l'étude des risques existentiels (CSER, pour *Centre for the Study of Existential Risks*) de Cambridge au Royaume-Uni, ou encore l'Institut de Recherche des Machines Intelligentes (MIRI, Machine Intelligence Research Institut) de Berkeley aux Etats-Unis. Amazon, Apple, Google, Facebook, IBM, Microsoft ont créé de leur côté en 2016 le *Partnership on AI to benefit people and society*, espace conjoint de réflexion éthique[52].

En Europe

L'*European Data Protection Supervisor* (EDPS) a créé fin 2015 l'*Ethics Advisory Group*[53] sur l'impact des innovations numériques dans la société et l'économie. Le parlement européen a voté en février 2017 un texte d'orientation sur *le droit civil en robotique*[54]. Le document de base comporte un *Code conduite éthique pour les ingénieurs en robotique* et un *Code de déontologie pour les comités d'éthique de la recherche*. La commission européenne a organisé, selon sa stratégie de *Digital Single Market*, une consultation publique incluant les questions de transparence des moteurs de recherche ainsi que l'utilisation des données collectées entre autres sur les plate-formes, qui a donné lieu à un retour public en janvier 2016[55]. On peut citer quelques éléments qui en ressortent : la régulation en vigueur ne répond pas aux questions de responsabilité introduites par les technologies relatives au Big data et aux objets connectés ; la crainte sur la transparence des plate-formes ; les situations d'hégémonie commerciales et de dynamique de marché etc.

A la suite de l'initiative Franco-Allemande sur l'Économie Numérique[56], un groupe de travail autour de la normalisation dans le domaine du Big data a été constitué. Il est piloté par l'AFNOR/la DGE du côté français et leurs homologues allemands DIN/ BMWi. Parmi les axes retenus comme « *Best practices* » à développer, on trouve les méthodes éthiques et responsables de traitement et de gestion des données massives. Ces recommandations sont relayées par un groupe de travail de BDVA sur les normes et standard (*Big Data Value Association*, qui pilote avec la commission européenne un partenariat public privé -PPP sur les Big Data à hauteur de 2, 5 milliards d'euros).

En France

Le Conseil national du numérique a abordé les problèmes de responsabilité des plate-formes, et notamment les questions de neutralité, transparence, et loyauté dans son « Avis sur la Neutralité des plate-formes[57] » en 2014. Depuis, ces sujets sont abordés dans plusieurs de ses avis, par exemple, sur la santé, la fiscalité, ou en lien avec le projet de loi

51 http://www.nytimes.com/2016/09/02/technology/artificial-intelligence-ethics.html

52 https://www.partnershiponai.org/

53 https://secure.edps.europa.eu/EDPSWEB/edps/EDPS/Ethics

54 http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+REPORT+A8-2017-0005+0+DOC+PDF+V0//EN

55 https://ec.europa.eu/digital-single-market/news/first-brief-results-public-consultation-regulatory-environment-platforms-online-intermediaries

56 http://www.economie.gouv.fr/deuxieme-conference-numerique-franco-allemande-a-berlin

57 https://cnnumerique.fr/wp-content/uploads/2014/06/CNNum_Rapport_Neutralite_des_plateformes.pdf

pour une République numérique[58]. Le Conseil d'Etat dans son étude sur *Le numérique et les droits fondamentaux*[59] en 2014 pose la question sur les prédictions dont sont capables les algorithmes d'apprentissage machine en recommandant de « *mieux encadrer l'utilisation d'algorithmes prédictifs à l'égard des individus* ». La loi *République numérique* confie à la CNIL l'animation des questions éthiques et sociétales soulevées par le numérique : un débat national sur l'éthique des algorithmes a été lancé le 23 janvier 2017[60]. Récemment, le Conseil Général de l'Économie, missionné par le secrétariat d'état à l'économie numérique, a organisé une consultation d'experts sur la régulation des algorithmes de traitement des contenus. Des préconisations ont été formulées en vue de vérifier la conformité aux lois et règlements dont la détection de discrimination illicite. Ces travaux ont débouché sur la création d'une plate-forme scientifique collaborative nationale « TransAlgo »[61] pour le développement de la transparence et de la responsabilité des algorithmes et des données. Le CNNum a rejoint l'initiative « TransAlgo » dans sa démarche nationale d'évaluation des plate-formes qui a fait l'objet de sa mission début décembre 2016[62] .

Enfin, au printemps 2017, l'Office Parlementaire d'Évaluation des Choix Scientifiques et Technologiques (OPECST) a publié une étude "Pour une intelligence artificielle maîtrisée, utile et démystifiée" [63] contenant 15 propositions, parmi lesquelles :

- *Proposition n° 2 : Favoriser des algorithmes et des robots sûrs, transparents et justes et prévoir une charte de l'intelligence artificielle et de la robotique.*
- *Proposition n° 3 : Former à l'éthique de l'intelligence artificielle et de la robotique dans certains cursus spécialisés de l'enseignement supérieur.*
- *Proposition n° 4 : Confier à un institut national de l'éthique de l'intelligence artificielle et de la robotique un rôle d'animation du débat public sur les principes éthiques qui doivent encadrer ces technologies.*

Le gouvernement a élaboré « Une stratégie pour la France en matière d'intelligence artificielle », FranceIA[64], intégrant également la dimension éthique.

VI. Conclusion

Les institutions et les citoyens prennent pleinement conscience de l'importance des questionnements éthiques posés par le numérique, et de leur diversité au delà du traitement des données à caractère personnel. L'effervescence autour de l'intelligence artificielle, et notamment de l'apprentissage machine, se concrétise par de nombreuses initiatives industrielles et de recherche à l'international, en Europe et en France, avec la particularité que la dimension éthique y est omniprésente. Le rôle des chercheurs est également de veiller à la qualité des plate-formes et des logiciels d'apprentissage machine en accès ouvert et aux préconisations de bon usage **[GEN-10]**.

58 https://cnnumerique.fr/plateformes/
59 http://www.ladocumentationfrancaise.fr/rapports-publics/144000541/
60 https://www.cnil.fr/fr/ethique-et-numerique-les-algorithmes-en-debat-0
61 http://www.economie.gouv.fr/files/files/PDF/Inria_Plateforme_TransAlgo2016-12vf.pdf
62 http://www.economie.gouv.fr/cge/modalites-regulation-des-algorithmes-traitement-des-contenus
63 http://www.senat.fr/presse/cp20170329.html
64 http://www.economie.gouv.fr/France-IA-intelligence-artificielle

Cette dynamique est propice à une **Initiative nationale fédérative de recherche sur l'impact sociétal et éthique des sciences et technologies du numérique [GEN-11]** afin de

- créer des synergies pour capitaliser sur les différents travaux et les enrichir ;
- encourager le dialogue entre la recherche et la société ;
- construire une parole française suffisamment forte pour entraîner une dynamique européenne ;
- émettre des préconisations de formation à tous niveaux ;
- valoriser l'engagement des chercheurs dans ces directions interdisciplinaires.

La structuration pourrait reposer sur un réseau d'animation, associant à parts égales des spécialistes des sciences et technologies du numérique et des spécialistes des sciences humaines et sociales.

Points d'attention et préconisations

*Par ailleurs, les préconisations générales **[GEN-x]** de la CERNA sur l'organisation de la recherche pour une meilleure prise en compte des questions éthiques en sciences et technologies du numérique, formulées en 2014, valent plus que jamais et sont rappelées ci-après.*

[GEN-10] Veille des chercheurs sur la qualité des plate-formes et des logiciels d'apprentissage machine en accès ouvert
Le chercheur contribuera à la veille sur la qualité des plate-formes et des logiciels d'apprentissage machine mis à la disposition du public, et la sensibilisation aux risques de mise en œuvre non maîtrisée à travers certaines applications.

[GEN-11] Initiative Fédérative de Recherche Numérique, Éthique et Société
Un réseau national de recherche pluridisciplinaire sur l'impact sociétal et éthique des sciences et technologies du numérique doit être créé afin de capitaliser de façon pérenne sur les diverses initiatives actuelles et de faire émerger un « positionnement français » susceptible d'impulser un dynamique européenne.

VII. Liste des préconisations

Rappel des préconisations générales de la CERNA

[GEN-1] Expertise et expression d'opinion
Lorsque le chercheur s'exprime en public sur une question de société relative à son activité professionnelle, il doit distinguer son intervention experte de l'expression de son opinion personnelle.

[GEN-2] Comités d'éthique opérationnels d'établissements
Il est recommandé que les établissements se dotent de comités opérationnels d'éthique en sciences et technologies du numérique.

[GEN-3] Initiatives des établissements sur les aspects juridiques
Il est recommandé que les établissements et autres acteurs concernés mettent en place des groupes de travail et des projets de recherche interdisciplinaires ouverts à l'international

et incluant des chercheurs et des juristes pour traiter des aspects juridiques des usages de la robotique.

[GEN-4] Sensibilisation et soutien du chercheur par les établissements
Il est recommandé que les établissements et autres acteurs concernés mettent en place des actions de sensibilisation et de soutien auprès des chercheurs et des laboratoires de recherche dans le numérique. Lors de l'élaboration et dans la conduite de ses projets le chercheur saisira, si nécessaire, le comité opérationnel d'éthique de son établissement.

[GEN-5] Données personnelles
Lors de la conception d'un système numérique ayant la capacité de capter des données personnelles, le chercheur se demandera si ce système peut être équipé de dispositifs facilitant le contrôle de sa conformité à la réglementation lors de sa mise en usage.

[GEN-6] Prévention d'attaque des systèmes numériques
Le chercheur veillera à prendre en compte l'exposition potentielle de ses recherches et prototypes à des attaques numériques malicieuses.

[GEN-7] Conduite de projet
Si le chercheur considère que le projet vise un développement pouvant avoir un impact important sur la vie des utilisateurs, il veillera à délibérer dès la conception du projet avec les acteurs et les utilisateurs potentiels afin d'éclairer au mieux les choix scientifiques et technologiques.

[GEN-8] Documentation
Le chercheur veillera à documenter l'objet ou le système conçu et à en exposer les capacités et les limites. Il sera attentif aux retours d'expérience à tous les niveaux, du développeur à l'utilisateur.

[GEN-9] Communication publique
Le chercheur veillera à faire une communication mesurée et pédagogique sachant que les capacités des objets et systèmes qu'il conçoit peuvent susciter des questionnements et des interprétations hâtives dans l'opinion publique.

[GEN-10] Veille des chercheurs sur la qualité des plate-formes et des logiciels d'apprentissage machine en accès ouvert
Le chercheur contribuera à la veille sur la qualité des plate-formes et des logiciels d'apprentissage machine mis à la disposition du public, et la sensibilisation aux risques de mise en œuvre non maîtrisée à travers certaines applications.

[GEN-11] Initiative Fédérative de Recherche Numérique, Éthique et Société
Un réseau national de recherche pluridisciplinaire sur l'impact sociétal et éthique des sciences et technologies du numérique doit être créé afin de capitaliser de façon pérenne sur les diverses initiatives actuelles et de faire émerger un « positionnement français » susceptible d'impulser un dynamique européenne.

Préconisations en éthique de la recherche en apprentissage machine

Dans l'ordre de leur formulation :

1-[DON-1] Qualité des données d'apprentissage

Le concepteur et l'entraîneur veilleront à la qualité des données d'apprentissage et des conditions de leur captation tout au long du fonctionnement du système. Les entraîneurs du système informatique sont responsables de la présence ou de l'absence de biais dans les données utilisées dans l'apprentissage, en particulier l'apprentissage « en continu », c'est-à-dire en cours d'utilisation du système. Pour vérifier l'absence de biais, ils doivent s'appuyer sur des outils de mesure qui restent encore à développer.

2-[DON-2] Les données comme miroir de la diversité

Les entraîneurs des systèmes d'apprentissage automatique doivent opérer le choix des données en veillant à ce que celles-ci respectent la diversité des cultures ou des groupes d'utilisateurs de ces systèmes.

3-[DON-3] Variables dont les données comportent un risque de discrimination

Les entraîneurs (qui peuvent être aussi les concepteurs ou les utilisateurs) doivent se poser la question des variables qui peuvent être socialement discriminantes. Il convient de ne pas mémoriser ni de régénérer par programmation ces variables, par exemple l'ethnie, le sexe ou l'âge. La protection des données à caractère personnel doit également être respectée conformément à la législation en vigueur.

4-[DON-4] Traces

Le chercheur doit veiller à la traçabilité de l'apprentissage machine et prévoir des protocoles à cet effet. Les traces sont elles-mêmes des données qui doivent à ce titre faire l'objet d'une attention sur le plan éthique.

5- [AUT-1] Biais de caractérisation

Le chercheur veillera à ce que les capacités d'apprentissage d'un système informatique n'amènent pas l'utilisateur à croire que le système est dans un certain état lors de son fonctionnement alors qu'il est dans un autre état.

6- [AUT-2] Vigilance dans la communication

Dans sa communication sur l'autonomie des systèmes apprenants relativement aux humains, le chercheur doit viser à expliquer le comportement du système sans donner prise à des interprétations ou des médiatisations irrationnelles.

7-[EXP-1] Explicabilité

Le chercheur doit s'interroger sur la non-interprétabilité ou le manque d'explicabilité des actions d'un système informatique apprenant. Le compromis entre performance et explicabilité doit être apprécié en fonction de l'usage et doit être explicité dans la documentation à l'usage de l'entraîneur et de l'utilisateur.

8-[EXP-2] Les heuristiques d'explication

Dans sa recherche d'une meilleure explicabilité d'un système apprenant, le chercheur veillera à décrire les limitations de ses heuristiques d'explication et à ce que les interprétations de leurs résultats ces soient exemptes de biais.

9-[EXP-3] Elaboration des normes

Le chercheur veillera à contribuer aux débats de société et à l'élaboration de normes et de protocoles d'évaluation qui accompagnent le déploiement de l'apprentissage machine. Quand les données concernent certains secteurs professionnels spécialisés (médecine, droit, transports, énergie, etc.), les chercheurs de ces domaines devront être sollicités.

10-[DEC-1] Place de l'humain dans les décisions assistées par des machines apprenantes

Le chercheur, concepteur de machines apprenantes d'aide à la décision, doit veiller à ce qu'aucun biais ne produise un résultat qui devienne automatiquement une décision alors que l'intervention humaine était prévue par la spécification du système, et prendra garde aux risques de dépendance de l'humain aux décisions de la machine.

11-[DEC-2] Place de l'humain dans l'explication des décisions assistées par des machines apprenantes

Le chercheur doit veiller à ce que les résultats du système soient autant que possible interprétables et explicables pour les personnes concernées, s'impliquer dans les nécessaires évolutions des métiers en faisant usage et dans l'encadrement des pratiques. Des agents experts contribueront à l'explication et à la vérification des comportements du système.

12-[CON-1] Choix des usagers d'activer ou non les capacités d'apprentissage d'un système

Le chercheur doit créer la possibilité d'utilisation de systèmes avec ou sans leur capacité d'apprentissage. Il doit donner à l'utilisateur au moins un paramètre de contrôle global sur la source des données utilisées pour l'apprentissage.

13-[CON-2] Consentement dans le cadre de projet

Le chercheur doit délibérer dès la conception de son projet avec les personnes ou les groupes identifiés comme pouvant être influencés.

14-[CON-3] Consentement à utiliser une machine capable d'apprentissage en continu

Le chercheur doit être conscient que la capacité d'apprentissage et la mise en réseau de telles capacités peut induire des problématiques nouvelles concernant autant le consentement de l'usager que celui de la société.

15- [RES-1] Mécanismes de contrôle

Le chercheur doit adapter et inclure dans le système, des mécanismes de contrôle, automatiques ou supervisés (par la machine ou l'être humain), sur les données, sur le fonctionnement au niveau informatique et sur le raisonnement suivi, afin de faciliter l'attribution de responsabilités au regard du bon fonctionnement ou du dysfonctionnement du système.

16- [RES-2] Déclaration des intentions d'usage

En documentant un système informatique apprenant, le chercheur doit décrire de manière sincère, loyale et complète, les limites qui lui sont connues de l'imputabilité d'une décision ou d'une action du système soit au code-source soit au processus d'apprentissage. Cette documentation servira de déclaration d'intention de la part du concepteur quant aux usages du système informatique qu'il envisage comme normaux. L'absence de telle déclaration ou son caractère tardif peuvent engager la responsabilité supplémentaire du concepteur.

ANNEXES

Présentation d'Allistene

En favorisant la recherche et les innovations dans le domaine du numérique, Allistene, l'alliance des sciences et technologies du numérique, accompagne les mutations économiques et sociales liées à la diffusion des technologies numériques. L'alliance a pour but d'assurer une coordination des différents acteurs de la recherche dans les sciences et technologies du numérique, afin d'élaborer un programme cohérent et ambitieux de recherche et de développement technologique. Elle permet d'identifier des priorités scientifiques et technologiques communes et de renforcer les partenariats entre les opérateurs publics (universités, écoles, instituts), tout en créant de nouvelles synergies avec les entreprises. Créée en décembre 2009, Allistene regroupe en tant que membres fondateurs la CDEFI, le CEA, le CNRS, la CPU, Inria et l'Institut Mines Télécom. Ses membres associés sont l'INRA, l'INRETS et l'ONERA.

Ses objectifs sont :

- Coordonner les acteurs de la fonction programmatique autour de priorités scientifiques et technologiques ;
- Élaborer des programmes nationaux répondant à ces priorités et des modalités pour la mise en œuvre de ces programmes ;
- Renforcer les partenariats et les synergies entre l'ensemble des opérateurs de la recherche du domaine, universités, écoles, instituts, et aussi les entreprises en particulier au sein des pôles de compétitivité du numérique ;
- Prolonger les priorités et programmes nationaux dans les différentes initiatives européennes et internationales relevant du domaine.

Site internet : www.allistene.fr

Présentation de la CERNA

La Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique a été créée fin 2012 par l'alliance Allistene.

Ses objectifs sont :

- Répondre aux questions d'ordre éthique posées par le Comité de Coordination d'Allistene ou par l'un des organismes membres ;
- Mener une réflexion sur l'éthique des recherches scientifiques développées en Sciences et Technologies du Numérique ;
- Sensibiliser les chercheurs à la dimension éthique de leurs travaux ;
- Aider à exprimer les besoins spécifiques de la recherche au regard du législateur et à les couvrir dans une démarche responsable ;
- Apporter un éclairage de nature scientifique aux décideurs et à la société sur les conséquences potentielles de résultats de recherche ;
- Veiller à ce que les étudiants soient formés sur ces questions ;
- Suggérer des thèmes de recherche permettant :
    - d'approfondir la réflexion éthique dans un cadre interdisciplinaire ;
    - de prendre en compte le résultat des réflexions éthiques.

Ses avis sont consultatifs, ils peuvent être rendus publics sous le contrôle conjoint des présidents de la CERNA et d'Allistene, après consultation de l'alliance. Ils doivent traiter de questions générales et participer d'une analyse de fond pouvant rendre compte de la diversité des échanges et des opinions de ses membres, tout en dégageant clairement des conclusions.

La CERNA ne traite pas les questions opérationnelles d'éthique et de déontologie, qui relèvent de la responsabilité des acteurs et de leurs établissements.

Site internet : http://cerna-ethics-allistene.org/