Big Data, algorithmes et marketing : rendre des comptes

Christophe BENAVENT
Professeur à l'Université Paris Nanterre

Cet article s'intéresse à la question de la mise en œuvre à vaste échelle d'algorithmes utiles au marketing, et s'intégrant dans une logique de plateforme. Prenant en compte des observations répétées d'externalités négatives produites par les algorithmes : ségrégation, biais de sélection, polarisation, hétérogénéisation, mais aussi leurs faiblesses intrinsèques résultant de la dette technique, de dépendances des données, et du contexte adversial dans lequel ils s'exercent, nous aboutissons à la nécessité d'une redevabilité algorithmique et nous nous questionnons sur la manière dont les algorithmes doivent être gouvernés.

Le Big Data pour le marketing n'est pas simplement un élargissement du spectre des méthodes d'étude, ni le déploiement à vaste échelle du recueil des données et l'exploitation d'une grande variété de formats d'information. C'est une intégration de l'information à la décision si intime que de la mesure à l'action il n'y a quasiment plus d'espace donné à la délibération.

Ce qui est désormais géré est un flux de décisions continu et contenu par l'architecture des algorithmes. C'est une transformation de la nature même du cycle de décision marketing (Salerno et al, 2013) qui pose un problème de nature éthique et politique : que se passe-t-il dans les boîtes noires de la société (Pasquale, 2015) et particulièrement celles des dispositifs marketing ? Comment en rendre compte ?

Deux événements en témoignent. Le cas de Volkswagen et de la tromperie algorithmique qui permettait d'échapper aux tests de pollution, ne trouble pas tant par le constat de la volonté stratégique de tricher et de cacher, que par l'hypothèse qu'il pourrait être un patch apporté et oublié par des ingénieurs incapables de résoudre la contrainte des législations environnementales. Le cas de Tay, l'agent conversationnel de Microsoft perverti par des activistes d'extrême droite qui l'ont entraîné à formuler un discours raciste, sexiste et suprémaciste, illustre une faille possible des systèmes auto-apprenants : ils sont dépendants des données.

Le premier cas relève certainement de ce que les informaticiens appellent la dette technique. Celui de Tay relève de l'informatique « adversiale ». Ces deux notions clés permettent d'éclairer ce qui fait la faiblesse des algorithmes. Cette faiblesse peut venir aussi de la nature du recueil de données, massif et intrusif, qui conduit à une réaction stratégique des sujets dont le mensonge est une des formes. Il est d'ailleurs une préoccupation importante des spécialistes de management des systèmes d'information comme en témoigne le travail de Joey George (2008). Le mensonge n'est pas une nouveauté, ni l'omission ; si la stratégie de leurre comme le

promeuvent Brunton et Nissembaum (2015) sert à protéger la vie privée, et c'est bien pour cela que les pièces comptables ont été inventées, les comptes doivent être prouvés. Les Sumériens ont inventé la comptabilité en même temps que les tribunaux. Pas de royaume sans scribe scrupuleux, ni recensement, ni mémoire ni registre[1]. La vitesse de calcul et son échelle sont nouvelles, pas le principe de compter ni celui de d'assurer l'intégrité des comptes.

Dans un monde de plateformes où les données et les algorithmes façonnent les services délivrés et les conditions de leur livraison, le regard critique sur le Big Data ne doit pas se tenir au registre de la dénonciation, il doit envisager les processus par lesquelles les algorithmes produisent des effets qui diffèrent de ce qui est attendu et des externalités négatives dans l'environnement où ils agissent. Il conduit à faire de l'idée de redevabilité algorithmique un concept central de la mise en œuvre des méthodes de Big Data

Gouvernementalité algorithmique : une nouvelle politique marketing

L'idée que les algorithmes ne sont pas neutres et agissent sur nos conduites se développe depuis plusieurs années. Lessig (2002) sans doute est un des premiers à voir et à vouloir mettre de la politique dans les algorithmes avec sa formule « code is law ». C'est cette idée qui est au centre du travail de Rouvroy et Berns (2013) ou de celui de Dominique Cardon (2015) qui en propose une sociologie perspectiviste.

L'idée de gouvernementalité, terme proposé par Michel Foucault[2], se définit comme l'action qui agit sur les conduites individuelles par des dispositifs de connaissance et de contrainte, pour gouverner la population et la ressource qu'elle représente. L'originalité de cette conception réside certainement en ce qu'elle dissocie l'objet du gouvernement, la population et son lieu d'exercice, la psychologie des individus. Dans un univers de plateformes qui est celui des moteurs de recherche, des marketplaces, des réseaux sociaux, cette gouvernementalité peut être saisie au travers de trois grands types de dispositifs et de ce qui propage leurs actions dans la population - les algorithmes - (Benavent, 2016).

Le dispositif principal est celui qui règle la capacitation des sujets de la population et qui, par son architecture et ses interfaces, leur permet d'agir tout en définissant les limites de cette action (restriction). Il peut se différencier selon les populations, et évolue en fonction de leurs comportements et de leurs interactions. Sur cette architecture se greffent les « *policies* », ou règlements intérieurs qui régissent les droits relatifs aux données personnelles, à la propriété sur les contenus et limite la liberté d'expression. Le troisième type de dispositif a pour finalité de motiver l'action comme les mécanismes de fidélisation, les indicateurs de réputation, la conception de nudges, le design d'un système de gamification destiné à motiver l'action des utilisateurs[3]. C'est le terrain de jeux des technologies persuasives dont Fogg est un des principaux promoteurs (2002)

1. C'est la raison d'être de l'effervescence aujourd'hui des débats sur la Blockchain.
2. Michel Foucault (2004), Sécurité, territoire, population, éditions du Seuil, 2004, « Par gouvernementalité, j'entends l'ensemble constitué par les institutions, les procédures, analyses et réflexions, les calculs et les tactiques qui permettent d'exercer cette forme bien spécifique, quoique très complexe de pouvoir qui a pour cible principale la population, pour forme majeure de savoir l'économie politique, pour instrument essentiel les dispositifs de sécurité » pp.111-112.
3. Les nudges sont des dispositifs qui prennent avantage des biais cognitifs pour orienter, sans imposer, la réponse des sujets dans une direction favorable au bien-être du sujet et de l'environnement. La gamification est cette discipline nouvelle née dans l'industrie des jeux vidéo, qui en utilisant les notes, les statuts, des scores de performance « ludifie» les activités apportant une gratification à ce dont l'utilité n'est pas pleinement perçue ou qui réclame un effort trop élevé, maintenant ainsi un niveau élevé de motivation.

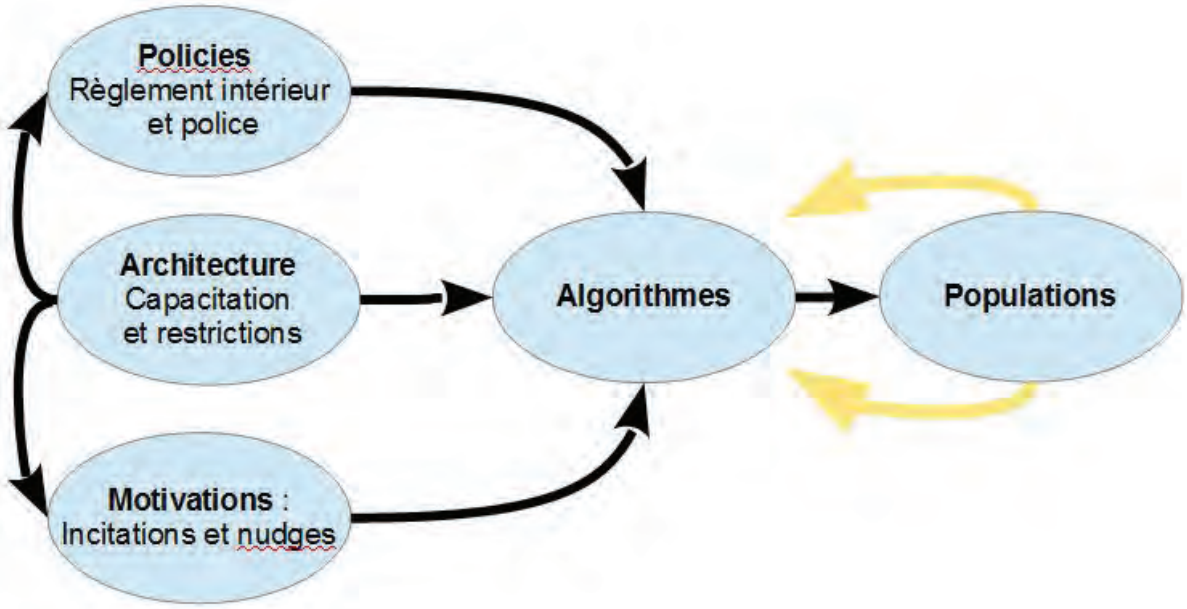


Figure 1 : les composantes de la gouvernementalité algorithmique

Les algorithmes ont un rôle d'abord évidemment de calcul dans l'instanciation du système complexe d'actions, de règles et de motivations, pour des situations particulières. Le système technique sur lequel ils s'appuient se caractérise par son volume (Big Data) et par sa granularité : l'échelle est celle de la minute et de quelques mètres. Ils peuvent être simples (un tri par ordre de prix) ou sophistiqués (le calcul d'un indice de sentiment à partir d'une méthode de machine-learning appliquée à des données textuelles). Mais plutôt que de mettre l'accent sur le calcul, c'est le rôle médiateur, la synchronicité du système qui en font un media équivoque.

Si on les examine sur un plan plus concret, comment les algorithmes exercent-ils leurs effets ? On s'aperçoit que pour le marketing, ils se concentrent sur un petit nombre de tâches. Ce à quoi servent les algorithmes s'inscrit dans la liste suivante :

- Filtrer et chercher : c'est bien le sens du Page Rank, mais aussi de très nombreuses méthodes, qui s'appuient sur l'historique des demandes dont on a montré depuis longtemps que sous-certaines conditions elles pouvaient présenter des résultats peu pertinents.
- Trier les fils de nouvelles : c'est l'outil privilégié par Facebook et qui fait l'objet d'une polémique récente à cause de l'obscurité de ses choix implicites.
- Recommander des profils, des produits : c'est le domaine d'application le plus important et un des plus anciens ; Amazon et Netflix en sont les porte-drapeaux. La question de la sérendipité en est un enjeu essentiel pour aller au-delà de recommandations trop évidentes.
- Indiquer des tendances et prédire des évolutions : une des armes de Twitter pour mettre en avant le contenu (trending topics). De manière plus générale c'est la prédiction des séries chronologiques : ventes, audience, à un niveau de plus en plus micro-économique.
- Calculer des scores de réputation, de risque, de qualité à des échelles jamais connues comme l'estimation des valeurs des maisons de Zillow (*Zestimate*), qui s'appuie sur un recensement de 185 millions de logements avec des milliers de caractéristiques.
- Catégoriser des images : le cas de Flickr, une plateforme de photos, est un très bon exemple d'application de deep learning pour taguer les images et ainsi améliorer leur « recherchabilité » et donc leur valorisation sur la marketplace.
- Produire et déclencher des alertes et notifications. C'est un enjeu du monde des objets connectés tels que les trackers, les balances connectées, les compteurs d'énergie, les

piluliers digitaux et la plupart des « applis » de nos smartphones.
Pour examiner avec plus de précision cette notion générale nous examinerons d'abord les effets sociaux indésirables que les algorithmes produisent : des effets de ségrégation et de biais de sélection, des effets de performativité, et au travers de certaines tentatives de correction, un accroissement potentiel de perte d'éléments de vie privée.

Ségrégation, polarisation et biais de sélection

Depuis quelques d'années de nombreux universitaires (les politiques sont bien moins nombreux) s'inquiètent du caractère obscur et ésotérique des algorithmes dont le fonctionnement peut même, à l'insu de leur concepteurs, produire des effets non-anticipés sur les populations. Tarleton Gillespie and Nick Seaver en maintiennent une bibliographie très dense sur *Social Media Collective*, un blog de recherche de *Microsoft*[4].

Un bon exemple de ce type d'effet est donné par l'étude de Eldelman et Luca (2014) sur les prix de location *Airbnb* à New-York. Ils observent une différence de 12$ entre les logements proposés par les Blancs et les Noirs. Ceci résulte principalement d'un mécanisme de réputation : l'exposition de la photo de l'hôte. L'algorithme est simplissime, il réside simplement dans le protocole de production d'une page d'offre. Ce n'est dans ce cas pas tant l'algorithme lui-même qui ségrégue, mais en « enactant » les décisions par la mise en évidence de la couleur de peau, il donne à cette dernière la valeur d'un attribut signalant un certain risque, conduisant ceux qui en sont victimes à s'ajuster en proposant des prix significativement plus bas.

Cet effet de ségrégation peut être généralisé à la notion de polarisation qui est au cœur de l'ouvrage : « The bubble society » (Pariser, 2011) mais semble aussi être une hypothèse forte sur la vie politique américaine. Comme l'indique un rapport de Pew Research (2014) même si le rôle des réseaux n'est pas démontré, on sait au moins que l'exposition aux opinions contraires ne concerne qu'un tiers des internautes. Il reste aussi à faire la part de l'effet purement algorithmique et de l'auto-sélection des individus (Bakshy et al,2015). Les réseaux sociaux amplifient-ils les liens homophiles en favorisant des phénomènes d'attachement préférentiels, ou ne sont-ils que le miroir de comportements soumis aux biais de confirmation ?

Les effets de biais de sélection peuvent s'observer sur l'agrégation des informations dans les réseaux sociaux. Un exemple simple illustre le phénomène : supposons que les individus qui postent le plus souvent, postent du contenu positif tandis que ceux qui postent moins souvent sont animés par la vengeance et seront donc négatifs. Supposons que ces deux groupes soient d'effectifs égaux et que les premiers postent 10 fois plus, la production collective de commentaires est composée à 80% de contenus positifs alors que la population qui les émet ne représente que 50%.

L'effet algorithmique est dans ce cas très simple : il résulte simplement d'une opération d'agrégation, mais il peut être volontairement amplifié comme l'a tenté Facebook dans une expérience contestée visant à filtrer les messages sur le fil de nouvelles des individus, en ôtant aléatoirement de 5% à 90% des contenus positifs et négatifs dans une population de 670 000 personnes. Le défaut éthique de l'expérience est que les sujets n'ont pas été informés conduisant les éditeurs à faire précéder la publication d'un avertissement. (Adam et Al, 2014). Le résultat de l'expérience est de démontrer que l'on peut propager l'émotion sociale en manipulant la composition du fil de nouvelles. Filtrer en renforçant les contenus positifs conduit les sujets à produire plus de posts positifs aussi.

4. Accessible sur https://socialmediacollective.org/reading-lists/critical-algorithm-studies/

Cet effet de sélection est amplifié par les retours d'information que constituent les notes et commentaires désormais employés de manière systématique. Après tout, chiffres et graphiques, alertes sonores ou ranking…tout élément qui indique une performance peut agir au moins par cette vieille idée des prophéties auto-réalisatrices. Il s'agit ici de la notion de performativité qui fait l'objet d'un regain d'intérêt dans les sciences de gestion (Berkowitz et Dumez, 2014). Elle propose qu'un acte de communication est aussi une action, le langage ne s'épuise pas en passant l'information de l'un à autre, il est aussi action sur ceux qui le reçoivent. Dans le domaine des études marketing, ce qui était réservé aux décideurs est désormais distribué à la foule, qui réagit en en connaissant le résultat. C'est ainsi que les systèmes de notation plutôt que de mesurer la satisfaction de manière fine et valide, présentent des distributions biaisées à droite (vers la satisfaction) et une variance faible : les usagers connaissant l'effet négatif sur la partie qu'ils évaluent (d'autant plus quant ils sont eux-mêmes évalués), préfèrent s'en tenir à une sorte de note de politesse.

Les développements récents en matière de Deep Learning font apparaître ce problème sous un visage différent. Il s'agit d'architectures de réseaux de neurones à plusieurs couches qui font l'objet d'un premier apprentissage non supervisé dont Yann Le Cun avec quelques autres chercheurs (Bengio et al. 2007) a relancé l'intérêt dans les années 2006-2007, et qui trouvent actuellement des terrains d'application importants dans le domaine de la reconnaissance d'objet dans les images, dans la reconnaissance vocale ou l'annotation de vidéos. Ils s'appuient sur des millions d'exemples qu'ils modélisent dans des espaces de plusieurs centaines de milliers, voire de millions de paramètres.

Le problème vient moins du calcul que de la manière dont sont présentés les objets dont on souhaite reconnaître des éléments de forme et pouvoir les associer à des catégories prédéterminées. Les catégories sont celles choisies pour l'algorithme par des humains. *Flickr* a entrepris de taguer automatiquement avec une centaine de catégories son stock considérable de 11 milliards d'images que les dépositaires répugnent à documenter par des mots clés. On choisit donc d'abord ce qu'il faut reconnaître[5]. Les ingénieurs de *Flickr* ont fait des choix, qui dans un second temps favoriseront certaines images plutôt que d'autre. S'ils définissent 12 éléments pour une catégorie « architecture », et seulement 4 pour décrire la catégorie « animaux », plus de détails sur l'une donne plus de chances à l'image d'être retrouvée, et ces choix sont redéfinis ensuite dans un processus d'apprentissage dont on ignore le protocole.

On retrouve ici une critique traditionnelle de la sociologie qui considère que les catégories produites indépendamment de l'effort théorique peuvent conduire à des contresens ; et l'on pourra reprendre par exemple l'analyse de Dominique Merllié (in Vatin, 2009) sur la discordance entre les déclarations de relations sexuelles des hommes et des femmes (1,57 contre 1,11 relations) où l'analyse élimine différentes explications (biais de désirabilité, exclusion des prostituées, effets de distributions) jusqu'à considérer le fait que par « relations sexuelles » hommes et femmes n'entendent pas la même chose. La constitution des catégories ne répond pas au critère de conventions qui s'établissent par la délibération. Dans le fil des travaux de Derosières et Thévenot (1988), il est désormais accepté que la statistique n'organise qu'un rapport de correspondance avec le réel, que les catégories qu'elle emploie sont le résultat de négociations, de conflits, de débats qui s'objectivent dans le consensus de la convention (Vatin, Caillé, Favereau, 2010). Sa validité tient à un accord sur ce qu'est la réalité.

Le problème dans la société des algorithmes est qu'ils ne font pas l'objet d'un tel débat. Ce monde proliférant de statistiques, est un monde sans convention ni accord, ou du moins seulement avec des accords partiels et peut-être partiaux.

5. Ceci est poussé au paroxysme avec les algorithmes « psychotiques » de l'équipe de Google Deep Dream qui les a entraîné à reconnaître des formes imaginaires dans des images ordinaires, générant des tableaux à la Jérôme Bosch.

Un algorithme juste

L'ignorance de ce type de phénomène peut conduire à des problèmes sociaux importants notamment lorsqu'on soupçonne que les algorithmes (même simples) produisent des effets de discrimination. Le cadre légal aux Etats unis a incité des chercheurs à s'intéresser à des méthodes de *fair scoring* qui tentent d'effacer l'aspect discriminant des algorithmes en minimisant le coût que représente la perte de précision L'algorithme juste n'est pas seulement celui qui prédit précisément, c'est celui qui produit un résultat socialement acceptable.

Prenons le cas des algorithmes de scoring courants dans le domaine bancaire pour attribuer ou non un crédit à la consommation. Ces algorithmes s'appuient sur ce que les spécialistes du machine learning appellent un classificateur, autrement dit une équation dont la forme très générale est la suivante : $S=f(X,\Theta)$

S, le risque, est le score calculé, (c'est mieux quand il s'exprime sous la forme d'une probabilité, celle de ne pas rembourser le prêt), en fonction d'un vecteur X de caractéristiques qui décrivent ce que l'on sait sur l'individu : âge, revenu, amis sur facebook, historique des mouvements bancaires, et peut être des données de santé. Θ désigne les paramètres du modèle. Ces modèles peuvent prendre une large variété de formes : modèle de régression, arbre de décision et random forest, SVM (Support Vector Machine), analyse discriminante, réseaux de neurones. Un tel modèle fournit, au mieux, une probabilité que le risque advienne.

En réalité il est mis en œuvre au travers d'une structure de décision. Elle peut être primitive quand on indique un seuil, par exemple : « si S est supérieur à 3% alors ne pas prêter ». Elle peut être un peu plus sophistiquée en pondérant gains et pertes espérées. Par exemple si G est la marge gagnée sur le prêt dans le cas où il n'y a pas d'incident, et P la perte subie si le client n'est pas en mesure de rembourser, le critère devient « $P*S+G*(1-S)>0$ ». On peut imaginer plus compliqué.

Le problème posé est qu'un tel algorithme n'est pas forcément juste, au sens de la précision. On connaît le problème classique des faux positifs. La théorie de la décision fournit des moyens de réduire, du point de l'entreprise, cet impact et de mieux définir sa stratégie (minimiser les risques, optimiser le gain...) ; mais du point de vue des individus qui, bien qu'en droit d'obtenir le prêt, ne l'auront pas, il y a injustice.

Il est assez facile de mesurer l'importance de ce risque simplement au moment où l'on teste le modèle. Il suffit de comparer les résultats à la réalité. On pourrait parfaitement exiger de ceux qui emploient de tels algorithmes, sans donner les paramètres Θ (qui sont un secret de fabrication), de rendre compte de la précision de leur algorithme ; et donc du risque de produire des décisions injustes. La plupart des modèles ne prédisent qu'assez mal les résultats, sauf dans les cas triviaux. Il faut garder en tête que même si la performance est remarquable, elle est loin d'être parfaite. Des jeux de données tests permettent d'en réaliser l'importance. Un exemple pour le machine learning est le jeu de données *Minist*, pour la reconnaissance de caractères ou le jeu de données CIFAR 100 pour la catégorisation des images. Il serait utile d'en disposer d'équivalents pour les applications marketing, auxquels les algorithmes commerciaux devraient se confronter ; les résultats de ces tests devraient être publiés.

Allons plus loin avec Dwork et al (2011). Un autre critère de justice est introduit, partant de la condition de Lipschitz qu'on peut exprimer assez simplement : un algorithme sera juste si la distance entre deux individus (telle qu'on la mesure au travers des caractéristiques X) est plus grande que la distance entre les scores calculés à partir de ces profils. C'est un critère de parité statistique à partir duquel les chercheurs proposent des modèles qui permettent de gommer les effets de variables discriminatoires. Cependant même si l'on introduit des approches justes

(fair) cela a un coût qui est celui de la confidentialité. Pour s'assurer que l'algorithme soit juste, il faut selon l'auteur que l'on détienne des données "sensibles", relevant de l'intimité. Un algorithme pour être juste devrait ainsi violer la vie privée. Ce qui amène un commentateur à relever qu'il n'est pas possible de construire un algorithme intrinsèquement juste, car même s'il est exact c'est au prix d'une perte de confidentialité. C'est à l'éthique de trancher : violer l'intimité ou éviter les discriminations.

Dette technique, dépendance des données et le risque adversial

Les effets inattendus des algorithmes sur les populations, le caractère peu maîtrisé des catégories qu'ils emploient, les difficultés de méthodes pour construire des algorithmes justes qui ne renforcent pas les différences qu'ils produisent ne sont pas les seuls problèmes, et s'ils résultent de l'interaction entre la technique et le social, d'autres sont propres à la technique. Trois notions proposées par les informaticiens sont utiles pour penser ces difficultés.

La première est la « dette technique » qui traduit qu'avec le temps les défauts des logiciels deviennent de plus en plus handicapants. Certains chercheurs, Sculley D. et al (2015), désignent ces effets inattendus par le terme de dette technique comme le fait une équipe de Google IA en décrivant par le menu les risques du machine learning : l'érosion des limites entre les composants stables du logiciel et les données dont leurs comportements dépendent ; l'intrication des paramètres et des données qui rend l'amélioration plus difficile que la mise en œuvre initiale ; les boucles de feed back cachées qui résultent de l'interaction du système et du monde ; les utilisateurs non déclarés ; la dépendance aux données et aux changements de l'environnement. Autant de problèmes qu'ils suggèrent de résoudre par des solutions graduelles et partielles. La dette technique est en quelque sorte le pendant des processus d'apprentissage qui graduellement accroissent la connaissance, c'est aussi l'accumulation des défauts, des rustines (*patch*), des restructurations incomplètes qui progressivement alourdissent le système. Un de ces principaux éléments résulte du phénomène de « dépendance aux données» qui désigne en informatique le fait qu'une instance utilise un résultat calculé précédemment, ce qui peut poser des problèmes quant à l'identification des erreurs dont il devient difficile d'isoler la source : des données inappropriées ou la structure même du modèle. Cette dépendance est forte quand on utilise des techniques à base de réseaux de neurones, et notamment leurs dernières générations qui possèdent de nombreuses couches et s'appuient sur des phases intermédiaires de filtrage. Les paramètres de ces algorithmes dépendent de données dont la production n'est pas forcément contrôlée et qui ne représentent pas tous les états naturels. L'effet se traduit par des modèles précis mais pas forcément stables dans le temps et qui évoluent au fil de l'évolution des populations.

Cette dépendance aux données est d'autant plus élevée que le « risque adversial » est important. C'est celui que fait peser l'individu qui veut déjouer les défenses et éventuellement exercer une nuisance. Les situations adverses sont celles dans lesquels le produit de l'algorithme peut être contré par un adversaire : par exemple le spammer qui veut échapper au filtre du spam. Certains résultats récents montrent ainsi qu'il est possible de modifier de manière imperceptible une image pour produire un mauvais classement/ inférence, et inversement qu'un algorithme peut apprendre à reconnaître ce qui n'est pas reconnaissable (comme l'expérience Google DeepMind l'illustre).

C'est en considérant cet aspect du problème qu'on comprend la grande différence entre le Big Data et la statistique traditionnelle telle qu'elle est pratiquée par les cabinets d'étude traditionnels. Cette dernière collecte des données sans affecter le corps social. Il y a juste des points de sonde, et elle les traite dans l'enceinte close du laboratoire de statistique, ses valeurs sont étudiées dans des salles de réunion. Dans la perspective du Big Data la collecte à grand échelle a toutes les chances de faire réagir le corps social qui répond moins sincèrement et plus stratégiquement.

Plus encore, la production et la diffusion de ses résultats étant instantanées c'est une seconde possibilité de réaction qui surgit.... et donc la nécessité de réajuster l'algorithme. L'évolution du moteur de recherche Google étudié par Dominique Cardon est un magnifique exemple. Les spécialistes du Search Engine Optimisation (SEO) ont appris à construire des univers web vides, sans autre contenu que celui qui est syndiqué, pour améliorer le référencement, conduisant Google à aménager ses critères pour en réduire le poids, mais alourdissant la conception même de l'algorithme. De ce point de vue les concepteurs et propagateurs d'algorithmes auraient intérêt à s'inspirer de la réflexion des statistiques publiques.

En revenant à la question des catégories que nous avons déjà évoquée, la perspective adversiale pose une question relative aux méthodes d'entraînement des algorithmes. Les choix extrêmes sont d'un côté celui d'un petit groupe d'experts qui définit les catégories et conclut sur le pronostic de la machine, ou, à l'opposé, le recours à la foule qui multiplie les épreuves, mais fait peser un risque d'entraînement inadéquat. Ce problème se rencontre dans la reconnaissance d'image, où l'apprentissage se fait dans le choix implicite de catégories auxquelles le réseau de neurones est confronté et tente de s'ajuster. Comment et pourquoi ainsi jugeons-nous la justesse de l'identification d'un lion dans une image ? Un félin, le symbole du lion, un lion proprement dit ? L'expert ou la foule par paresse ou superficialité risquent de confondre allégrement les lions représentés (des photographies de lions singuliers) et les représentations de lion (ceux des dessins animés). Il n'y aura pas eu de discussion pour convenir que la catégorie se définit comme « images de lion ». Les catégories ne correspondent pas tant à une réalité naturelle qu'à la réalité de nos catégories, et les machines n'élaborent pas encore les catégories même si elles peuvent apprendre à les reconnaître.

Dans la conception des algorithmes de cette espèce, ressurgit une tâche essentielle qui correspond à l'idée de validité de contenu, qui consiste à s'assurer que la catégorisation suit des protocoles particuliers qui assurent la légitimité des catégories et de la catégorisation. Ils sont probablement intermédiaires entre les experts et la foule et doivent présenter une qualité délibérative.

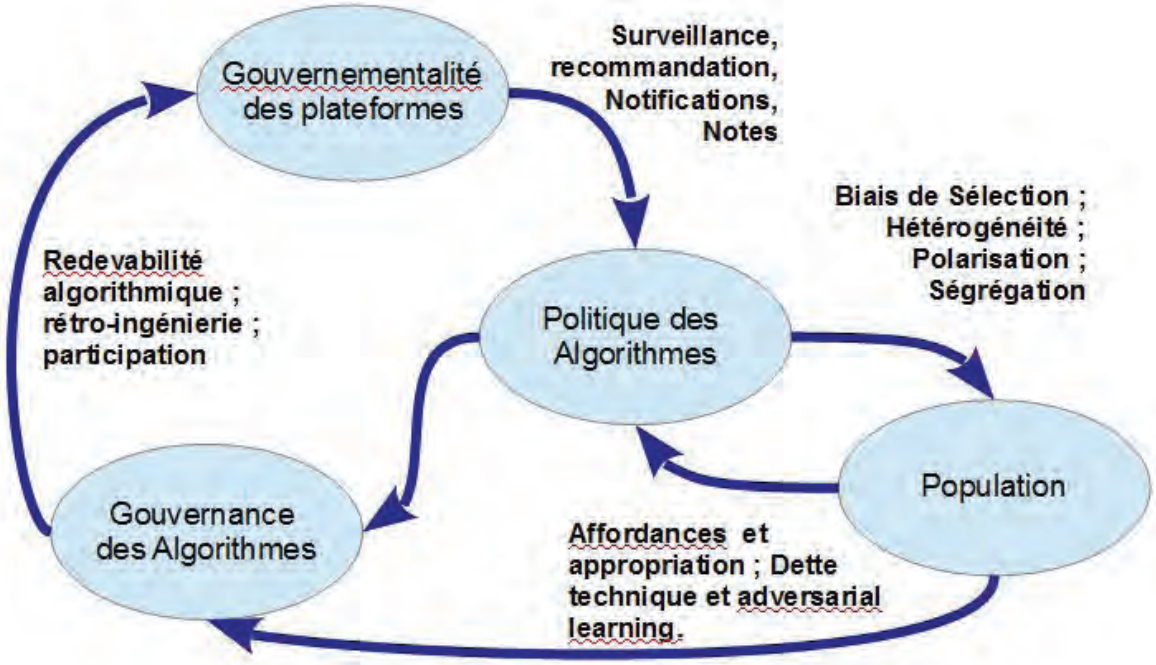
Les formes de la redevabilité algorithmique

L'ensemble de ces problèmes mobilise désormais les chercheurs, notamment en droit, autour de la notion d'*algorithmic accountability*[6]. Un vieux terme français « redevabilité » respecte mieux un esprit moins dominé par l'idée de l'agence et des défauts d'information inhérents, que par celle de l'intendance (théorie du *stewardship*) qui suggère que l'on peut être responsable aussi de ce que l'on ne possède pas : le bien public, le bien des autres, le royaume de Dieu... L'idée reste la même : celle d'une nécessité de rendre des comptes, de justifier des actions entreprises.

Si les algorithmes peuvent produire des effets inattendus et injustes, même s'ils ne sont pas le fruit d'une intention maligne, il est encore plus urgent de s'interroger sur la nécessité et les formes de leur obligation de rendre compte de leurs conséquences.

6. Comme l'a inauguré l' « Algorithms and Accountability Conference » tenue à la New York University le 28 février 2015.

De la gouvernmentalité des plateformes
à la gouvernance des algorithmes



Cette obligation se traduit aujourd'hui par la nécessité de déverrouiller les boites noires. C'est une exigence politique croissante, qui conduit à l'injonction que les plateformes doivent rendre compte des méthodes qu'elles emploient à la société dans son ensemble, en se cachant moins derrière la confidentialité des procédés.

C'est moins une question de droit que de politique, car c'est le politique qui aménage le droit. Cette exigence est d'autant plus naturelle que les algorithmes sont le plus souvent empruntés au domaine public comme le fut le page rank de Google ou la méthode de filtrage collaboratif d'Amazon.com. Mais ce n'est pas le problème principal. Par redevabilité on entend le fait que les effets de ces algorithmes doivent être considérés comme n'importe quelle externalité ; il est devenu évident aujourd'hui que dans les rapports d'activités, les effets environnementaux et sociaux de l'activité soient mentionnés.

On notera, avant d'aller plus loin, que dans la loi intitulée « Loi pour une République Numérique » (promulguée le 7 octobre 2016), cette dimension est relativement absente (sauf pour l'administration avec l'article 2).Si le droit à la vie privée semble être parfaitement reconnu et étendu (le droit de contrôle et de rectification va s'étendre à celui de la restitution des données sous une forme portable), rien par contre ne concerne ce que l'on fait de ces données, à l'exception des rapprochements autorisés ou d'éléments relatifs à la revente des données. Aucune exigence n'est formulée à l'égard des données transformées par les algorithmes, sauf un principe de loyauté des plateformes et de production d'indicateurs de transparence (titre II, section 3). Cette observation est un deuxième argument qui milite en faveur d'une obligation à rendre compte qui doit être imposée aux plateformes.

S'il faut formuler ces points d'obligation, la réflexion doit suivre leur nature. Un algorithme est une suite finie et non ambiguë d'instructions qui permettent d'obtenir un résultat ou de résoudre un problème. Il demande des entrées, il se caractérise par sa finitude, son exactitude et son rendement, et se traduit par ses sorties.

Les entrées soulèvent le premier problème. Ce problème est celui de la sincérité, de la fiabilité, de la validité et de la précision des informations qui sont introduites dans l'algorithme. Dans le monde qui est le nôtre ces entrées sont massives, entachées d'erreurs, de mensonges, d'omissions, souvent d'inexactitudes. Rendre compte de l'algorithme c'est donc très simplement préciser la qualité de ces informations entrantes et leur ôter l'évidence de leur véracité. Il s'agit aussi d'évaluer les effets de leur distribution incontrôlée sur les résultats.

Le traitement est un second problème. La finitude concerne notamment la nature des méthodes de calcul employées. À titre d'exemple dans les méthodes d'estimation d'algorithmes statistiques on sait qu'il y a des problèmes de minimum local et que, quel que soit le volume des données traitées, il n'est pas toujours assuré qu'une stabilité des paramètres soit obtenue. De manière plus sophistiquée il est indispensable que les algorithmes ne produisent pas dans leur décision des effets de faux positif ou du moins rendent compte de la balance de ces effets. Un bon exemple est celui de la lutte anti-terrorisme dont certains ont calculé les « effets secondaires » : pour identifier 3000 terroristes parmi 35 millions de personnes, un système précis (les identifiant à 99%) et faisant peu d'erreurs (accuser à tort 1% des innocents) générera près de 350 000 alertes, dont seules 2970 correspondront à de vrais « positifs ». Le rendement ici est particulièrement faible et l'injustice patente.

Quant aux sorties, il s'agit d'examiner leurs effets sociaux : compétition accrue, discrimination ou polarisation. L'obligation d'examiner ces effets pourrait s'inspirer des procédures imposées à l'industrie pharmaceutique pour limiter les effets secondaires, ou des obligations environnementales de tout un chacun.

La redevabilité algorithmique en ce sens s'approche d'une responsabilité particulière qui provient non pas d'un mandat que l'on a reçu d'un propriétaire, mais de celui qu'octroie la société dans son ensemble, celui d'un bien commun. Les algorithmes et le Big Data se nourrissent de ce qui est à la fois un commun (au sens où nulle propriété ne définit la valeur qu'ils produisent de manière agrégée) et de ce que leur mise en œuvre peut affecter l'environnement commun. Il reste à en définir les modalités, ce qu'il conviendra d'appeler gouvernance des algorithmes. Les GAFAs ne s'y trompent pas ! Conscients des problèmes de légitimité qui pourraient advenir, ils ont lancé le partenariat pour l'IA : « Partnership on Artificial Intelligence to Benefit People and Society ».

Conclusion

Les choses sont donc claires : ne comptons pas sur les machines pour réduire les problèmes sociaux, la responsabilité de ce que font les machines appartient entièrement aux humains, à leurs concepteurs et à leurs propriétaires.

Plutôt que d'interdire ou d'autoriser, la première étape de la régulation du monde des machines est d'imposer socialement l'exigence de rendre compte, ne serait-ce que par une approche d'ingénierie inversée comme le propose Diakopoulos (2014) qui vise à partir de l'étude des données d'entrée et de sortie à reconstituer le fonctionnement effectif des algorithmes, en démontant à rebours leur mécanique pour en retrouver les principes et identifier la source des effets pervers. Au-delà, la pression sociale et juridique va tendre à imposer aux gestionnaires des algorithmes la production d'études d'impacts, et dans certains cas, à engager des politiques compensatoires. C'est ainsi le cas de Airbnb en manière de discrimination qui se manifeste aujourd'hui par une invitation faite aux utilisateurs de signer une charte anti-discrimination.

La nécessité d'ouvrir les boites noires au moins partiellement s'impose. Les algorithmes doivent rendre des comptes et pas seulement sur leur efficacité. Quels biais de jugement véhiculent-ils ? Quels sont les risques de ségrégation et de discrimination ? Quelles injustices produisent-ils ?

Sur quelles conventions acceptables sont-ils conduits ?

La redevabilité algorithmique, s'impose moins comme solution à un conflit d'agents (des consommateurs qui prêtent leurs données en échange d'un service sans connaître ce qui en est fait) que comme responsabilité globale à l'égard de la société et de ses membres car, peu ou prou, les algorithmes façonnent l'univers où nous vivons. Cet univers est commun. On y démêle difficilement les contributions des uns et des autres. Dans un monde où les asymétries d'information sont majeures - certains contrôlent les algorithmes qui génèrent de la valeur et des millions d'autres n'en sont que les objets - on ne peut guère espérer que la distribution des incitations puisse renverser les choses. Les opérateurs prendront certainement soin de la légitimité des algorithmes, il n'est pas sûr qu'ils aillent plus loin. Il faut espérer au moins obtenir que les algorithmes n'œuvrent pas contre l'intérêt commun.

Pour les marketers, la principale conséquence est qu'il faudra prendre en compte dans la conception des algorithmes - ceux qui fixent des prix, ceux qui définissent un niveau d'assurance, ceux qui décident de l'octroi d'un crédit, ceux qui façonnent les environnements éditoriaux, ceux qui animent les places de marché - non seulement l'acceptation sociale de leurs technologies mais aussi les effets inattendus de leurs machines imparfaites. Ils ne devront pas ignorer la dimension politique de « leur machinerie ».

Références

- Adam et al. (2014)
- Bakshy, Eytian, Solomon Messing, Lada A. Adamic. (2015), "Exposure to ideologically diverse news and opinion on Facebook", Science. 2015 Jun 5;348(6239).
- Benavent Christophe (2016)
- Bengio Yoshua and Yann LeCun: Scaling learning algorithms towards AI, in Bottou, L. and Chapelle, O. and DeCoste, D. and Weston, J. (Eds), Large-Scale Kernel Machines, MIT Press, 2007,
- Berkowitz Héloïse, Dumez Hervé (2014) Un concept peut-il changer les choses ? Perspectives sur la performativité du management stratégique, Actes AIMS.
- Brunton, Finn et Helen Nissenbaum, Obfuscation : *A User's Guide for Privacy and Protest*, Cambridge, MIT Press, 2015,
- Cardon, Dominique (2015), *A quoi rêvent les algorithmes. Nos vies à l'heure des Big Data*, Seuil/La République des idées.
- Derosières, Alain et Laurent Thévenot (1988) « les catégories socio-professionnelles », La Découverte, Collection Repère, 1988.
- Diakopoulos, Nicholas (2014) Algorithmic-Accountability : the investigation of Black Boxes, Tow Center for Digital Journalism. June 2014.
- Dwork, Cynthia; Hardt, Moritz; Pitassi, Toniann; Reingold, Omer; Zemel, Rich (2011) « Fairness Through Awareness », arxiv, arXiv:1104.3913
- Edelman, Benjamin and Michael Luca (2014), « Digital Discrimination: The Case of Airbnb.com », Harvard Business School, Working Paper 14-054 January 10, 2014
- Fogg, B. J. (2002). Persuasive Technology: Using Computers to Change What We Think and Do. Morgan Kaufmann
- George, Joey F. and Robb, A. (2008) "Deception and Computer-Mediated Communication in Daily Life." Communication Reports 21(2), 2008, 92-103.
- Goodfellow,Ian J. , Yaroslav Bulatov ,Julian Ibarz , Sacha Arnoud, Vinay Shet(2014), Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks, (Submitted on 20 Dec 2013 (v1), last revised 14 Apr 2014 (this version, v4)), arXiv.org, arXiv:1312.6082
- Guérard, Stéphane, Ann Langley, and David Seidl 2013. "Rethinking the concept of performance in strategy research: towards a performativity perspective." M@N@Gement 16, no. 5: 566-578.
- Kramera, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock, (2014) « Experimental evidence of massive-scale emotional contagion through social networks », PNAS, vol. 111 no. 24
- Lessig (2002)
- Michel Foucault (2004), Sécurité, territoire, population, éditions du Seuil
- Pasquale, Frank (2014), *The Black Box Society*, Harvard University Press
- Pariser, E. 2011. *The Filter Bubble*. Penguin.
- Pew Research, (2014) « Political Polarization in the American Public » Pew Research, June 2014
- Rouvroy, Antoinette et Berns Thomas, (2013) « Gouvernementalité algorithmique et perspectives d'émancipation » Le disparate comme condition d'individuation par la relation ?, Réseaux, 2013/1 n° 177, p. 163-196.
- Salerno, Francis, Christophe Benavent, Pierre Volle, Delphine Manceau, Jean-François Trinquecoste, Eric Vernette et Elisabeth Tissier-Desbordes (2012), Eclairages sur le marketing de demain : prises de décisions, efficacité et légitimité, Décisions Marketing, Oct dec, n°72
- Sculley D. , Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young (2015), "Machine Learning: -The High-Interest Credit Card of Technical Debt", SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)
- Vatin François, Caillé Alain, Favereau Olivier, (2010) « Réflexions croisées sur la mesure et l'incertitude », Revue du MAUSS 1/2010 (n° 35) , p. 83-10
- Vatin, François (2009), Evaluer et Valoriser, Une sociologie économique de la mesure, Presse Universitaire du Mirail