2015

Université de Technologie
de Compiègne

Félix Motot

MEMOIRE BC01 :

Savoir et pouvoirs algorithmiques : enjeux du « Big Data »

Sommaire

Introduction

Le concept de « Big Data » est désormais largement répandu et semble familier à la plupart des individus. Pourtant, il suffit d'enquêter sur ce que « monsieur tout le monde » entend par ce terme pour s'apercevoir que sa réponse est plutôt vague et imprécise, et qu'il est souvent confondu avec d'autres notions comme le Data Mining. Kate Crawford, chercheur chez Microsoft dans le domaine des réseaux sociaux, et Jason Schultz, enseignant en droit à l'Université de New York, définissent d'ailleurs le Big Data comme un « terme général et imprécis » [Crawford - Schultz, 2014]. Nous allons donc dans un premier temps établir la définition de Big Data que nous utiliserons dans cet exposé, en nous basant sur celle fournie par Crawford et Schultz ; selon ces chercheurs, le Big Data est constitué de la croyance en une fiabilité des résultats corrélée positivement avec la quantité de données récoltées, à laquelle s'ajoute une confiance dans les algorithmes et le calcul informatique, en tant que meilleurs moyens de traiter ces données. Enfin, le Big Data fait aussi référence selon eux aux outils qui exploitent ces moyens de traitement.

« In practice, the term encompasses three aspects of data magnification and manipulation. First, it refers to technology that maximizes computational power and algorithmic accuracy. Second, it describes types of analyses that draw on a range of tools to clean and compare data. Third, it promotes the belief that large data sets generate results with greater truth, objectivity, and accuracy »[1]. [Crawford - Schultz, 2014, p.96]

Notre définition sera donc très proche de celle, précise et complète, énoncée ci-dessus : Le Big Data fait, selon nous, d'une part référence aux quantités massives et hétérogènes de données, en provenances de sources très diverses, qui sont désormais disponibles sur internet, et auxquelles doivent faire face différents acteurs; D'autre part aux nouvelles méthodes proposées pour faire face à ces données, les traiter et en sortir des informations exploitables.

Nous sommes confrontés, depuis l'arrivée d'internet, à une prolifération exponentielle des données, à tel point que selon IBM, 90% des données existantes dans le monde auraient été créées au cours des deux dernières années. Leur provenance est de plus en plus diverse (capteurs de toutes sortes, contenus publiés sur les réseaux sociaux, vidéos, données GPS, …), et la prolifération de ces contenus toujours plus hétérogènes minimise les capacités intellectuelles humaines, qui sont devenues très largement insuffisantes pour appréhender, traiter et recouper l'ensemble des données délivrées quotidiennement sur internet et dans les entreprises. Le Big Data est donc aujourd'hui un enjeu essentiel et stratégique, puisqu'il se fixe comme objectif premier de permettre la manipulation de ces très grandes quantités de données non structurées, en établissant des corrélations entre elles. Le forum économique

[1] « *en pratique, le terme englobe trois aspects de l'explosion du volume de données et de leur manipulation. En premier lieu, il se réfère aux technologies qui maximisent le pouvoir des traitements informatiques et la précision algorithmique. En second, il décrit des types d'analyses qui utilisent une sélection d'outils pour nettoyer et comparer les données. Enfin en troisième et dernier lieu, il promeut la croyance qu'une grande quantité de données génère des résultats plus précis, objectifs et par là plus vrais*. »

mondial[2] a d'ailleurs estimé, en 2014, qu'il constituait désormais une ressource comparable au pétrole sur le plan économique.

Il est donc omniprésent, car ces volumes de données de plus en plus importants et de moins en moins structurés présentent un fort enjeu commercial, marketing, et donc économique pour les entreprises, mais constituent également un potentiel important pour les gouvernements à des fins de sécurité et de contrôle des populations ; enfin elles intéressent les scientifiques pour la quantité de connaissances qu'elles peuvent représenter après traitement, sans oublier les individus eux-mêmes qui voient manifestement un intérêt à pouvoir partager leurs données instantanément à l'ensemble du globe, puisqu'ils produisent de plus en plus de contenus spontanément sur les réseaux sociaux et les autres plateformes d'échanges (blogs, …).

Il affecte toutes les strates de la vie des individus, et ce sans leur laisser le choix ni même leur poser la question; le Big Data modifie en effet les comportements sociaux, bouleverse profondément la science et la recherche, et va jusqu'à influencer les modes de gouvernance. La révolution qu'il a apporté dans la pratique des sciences sociales, qui jusqu'à présent reposait sur des observations et des analyses qualitatives, et qui désormais se basent de plus en plus sur le modèle computationnel[3] du Big Data, est révélatrice des bouleversements apportés par cette nouvelle ressource. Selon les promoteurs du Big Data, lorsque le volume de données est suffisamment important, les chiffres parlent d'eux-mêmes, et ces données n'ont plus besoin d'être analysées et interprétées. Paradoxalement, on s'aperçoit que l'augmentation de son impact sur nos vies n'a pas été suivie par les capacités d'appréhension du public, et parfois même ni par celles de ceux qui conçoivent les algorithmes ; Bernard Stiegler, philosophe français qui s'intéresse aux enjeux liés aux mutations actuelles portées par le développement technologique, rapporte en effet que « de plus en plus d'ingénieurs participent à des processus techniques dont ils ignorent le fonctionnement, mais qui ruinent le monde » [Stiegler, 2011]. Cela explique peut-être en partie pourquoi il est si peu légiféré, ni même encadré.

Comme le Big Data semble représenter un enjeu de plus en plus essentiel et omniprésent dans nos sociétés, il s'agira dans cet exposé de se demander dans quelle mesure il est présent dans la vie des individus et des sociétés, si cette présence est légitime, et le cas échéant si on peut s'en protéger et comment. Nous nous demanderons par ailleurs pourquoi, comme le relèvent à juste titre Crawford et Schultz, les algorithmes bénéficient de la croyance du public en leur objectivité et leur fiabilité, et si cette croyance est justifiée.

Dans un premier temps, nous tâcherons donc de démontrer que le Big Data est aujourd'hui très fortement présent dans la vie quotidienne des individus, et nous étudierons la relation entre l'individu et le Big Data. Cette première partie nous permettra également d'aborder

[2] *« le Forum économique mondial (World Economic Forum) est une fondation à but non lucratif dont le siège est à Genève. Le Forum est connu pour sa réunion annuelle à Davos, en Suisse, qui réunit des dirigeants d'entreprise, des responsables politiques du monde entier ainsi que des intellectuels et des journalistes, afin de débattre des problèmes les plus urgents de la planète* ». – Dictionnaire du commerce international

[3] Un modèle computationnel est un modèle qui utilise le traitement d'un ordinateur.

l'évolution des pratiques gouvernementales liées au Big Data, et les implications en termes de législation. Nous nous demanderons dans une seconde partie d'où vient la réputation d'objectivité des algorithmes, et si cette réputation est fondée. Enfin nous tâcherons, dans une troisième et dernière partie, de savoir si l'individu dispose de moyens pour lutter contre cette collecte massive des données personnelles.

Pour chacun de ces points, nous aborderons par ailleurs les questions éthiques qu'ils impliquent.

Relation entre Big Data et individus

Le Big Data fait partie intégrante de la vie des individus ; il a en effet joué un rôle dans l'évolution de nos pratiques, de façon progressive et imperceptible pour la plupart des individus, comme par exemple la façon dont nous cherchons et utilisons l'information, dont nous communiquons avec nos proches, ainsi que nos pratiques de consommation. Nous allons dans cette première partie nous concentrer sur l'explicitation de l'évolution de ces comportements, et nous montrerons également que cette influence n'est pas unilatérale, du Big Data vers l'individu, mais bel et bien bilatérale, car l'individu sait lui aussi, d'une part s'adapter au Big Data pour en tirer le meilleur profit, parfois sans même en avoir conscience, et d'autre part il oriente la façon dont ses algorithmes évoluent puisqu'il est le garant de leur acceptation.

Influence sur les comportements

Morozov, Weisberg, Pariser, Crawford, Gillespie

Les sites internet adaptent de plus en plus leurs services aux particularités de chaque utilisateur, aussi les résultats d'une même recherche sur Google effectuée par deux utilisateurs différents divergent de plus en plus, et ce même pour des requêtes extrêmement simples. Pour reprendre l'exemple des moteurs de recherche, ceux-ci cherchent à identifier le « profil » de chaque utilisateur afin de sélectionner les résultats intéressants pour lui, là où un moteur de recherche non personnalisé se contenterait de fournir des résultats pertinents. Ainsi, selon certains chercheurs de Microsoft, la personnalisation des résultats de recherche favoriserait la sérendipité[4], car elle maximiserait les chances de survenue de l'inattendu, en nous orientant vers les bonnes informations au bon moment, selon notre état d'esprit. Comme le relève Tarleton Gillespie, professeur au département communication à la *Cornell University*, les algorithmes jouent donc un rôle de plus en plus important dans la détermination des informations pertinentes pour les individus, en fonction de leurs préférences, de leur façon de s'exprimer ou encore des traces laissées par leur activité sur le web.

Les géants du web ne font pas que nous fournir des informations, ils nous fournissent également leurs algorithmes, qui sont modifiés à chaque nouvelle utilisation en fonction des dernières activités de l'utilisateur. Les résultats fournis par Google ne sont pas statiques, car ses algorithmes font évoluer les profils utilisateurs au fur et à mesure que ceux-ci utilisent ses services et lui fournissent des informations sur eux ; cela permet une adaptation continue des résultats qu'il fournit en fonction de ces nouvelles informations. Les acteurs du Big Data créent en quelque sorte une « identité algorithmique » [Cheney - Lippold, 2011] à chaque utilisateur ; ils cherchent à créer une copie virtuelle de chaque individu, en analysant ses traces de navigation et en cherchant à en connaître le plus possible sur lui. La prédictivité d'un individu est en effet d'autant plus grande que les données le concernant sont nombreuses ; comme le soulignent très justement Antoinette Rouvroy, chercheuse qualifiée du FNRS au

[4] « *La sérendipité est le fait de réaliser une découverte scientifique ou une invention technique de façon inattendue à la suite d'un concours de circonstances fortuit et très souvent dans le cadre d'une recherche concernant un autre sujet*. » - Wikipédia

centre de Recherche en information, droit et société, et Thomas Berns, chercheur en philosophie politique et morale, afin de représenter justement un individu et le rendre « calculable », les données doivent être « *à la hauteur du réel lui-même* » [Rouvroy – Berns, p172, 2013]. A cet effet, leur objectif est d'enfermer au maximum l'utilisateur dans leur propre écosystème. L'exemple de Google est frappant, car il l'entoure d'une multitude de services (réseau social Google+, Google translate, messagerie Gmail, vidéos Youtube, …), qui lui permettraient presque de se contenter uniquement des services de Google au cours de sa navigation. Ce phénomène va même encore plus loin grâce à ce qu'on appelle les interfaces de programmation applicatives (API[5]), qui permettent aux géants du web d'être présents sur des sites tiers, et donc de continuer à tracer les utilisateurs sur ces sites. Nous citerons l'exemple de Facebook et de ses « *like* » intégrés sur de nombreux sites internet n'appartenant pas au groupe. Ces gros acteurs du web accumulent donc de très grandes quantités d'informations sur les individus, et la prolifération des applications dites « sociales » sont d'autant plus bénéfiques qu'elles encouragent les individus à produire spontanément de plus en plus de contenus (statuts, publication de vidéos, …) ; la quantité d'informations dont ils disposent sur les individus va donc potentiellement s'accroître de plus en plus rapidement, et les individus les fourniront de plus en plus spontanément.

Cette idée de représentation numérique de l'individu dans sa singularité est toutefois à nuancer car d'une part, les acteurs du Big Data ne font pas qu'accumuler des informations hétérogènes et complexes à analyser, ils nous demandent de nous conformer à eux pour pouvoir ainsi mieux nous analyser selon leur propres critères : par exemple les règles de publication, d'interaction et de comportement sur les réseaux sociaux sont codifiées par l'entreprise qui le développe et les utilisateurs doivent s'y conformer. D'autre part, comme le soulignent Antoinette Rouvroy et Thomas Berns, le Big Data nous créé un double numérique, mais qui ne représente pas certaines dimensions de notre singularité, puisque les algorithmes se contentent d'établir des relations entre de très grandes quantités de données, et ne s'intéressent absolument pas à ce que nous sommes « en particulier » : « *il se contente de s'intéresser et de contrôler notre « double statistique », c'est-à-dire des croisements de corrélations, produits de manière automatisée, et sur la base de quantités massives de données, elles-mêmes constituées ou récoltées « par défaut ». Bref, ce que nous sommes « en gros », pour reprendre la citation d'Éric Schmidt, ce n'est justement plus aucunement nous-mêmes (êtres singuliers).* » [Rouvroy – Berns, p180, 2013]

Au-delà des problèmes que peuvent impliquer, notamment en termes de vie privée, le fait que tant d'informations sur les individus reposent entre les mains de quelques-uns, selon ses détracteurs, comme Eli Pariser, président du conseil de MoveOn[6], cette personnalisation serait

[5] Une « Application Programming Interface » est « *un ensemble normalisé de classes, de méthodes ou de fonctions qui sert de façade par laquelle un logiciel offre des services à d'autres logiciels.* » - Wikipedia

[6] « *MoveOn est une association politique basée aux États-Unis. Elle se décrit comme politiquement progressiste et ancrée à gauche.* » - Wikipédia

également néfaste car elle enfermerait l'individu dans un « *cocon informationnel* » [Pariser], dans lequel l'utilisateur n'est plus confronté aux informations qui lui sont désagréables, et ne voit donc plus dans le web qu'un reflet de ses propres valeurs et ses propres désirs. Pariser nous dit que tous ces efforts pour personnaliser les recherches sur internet et notre expérience sur les réseaux sociaux nous rend plus renfermés et égocentriques, car moins ouverts à ce qui nous surprend et ne nous attire pas spontanément. Pariser reprend d'ailleurs l'argument de la sérendipité pour le retourner, car selon lui la non-personnalisation augmente nos chances d'être confrontés à des informations inattendues, et par là l'exposition à de nouvelles idées. De même, la personnalisation nous rendrait plus vulnérables à la manipulation et à l'endoctrinement, car notre confiance dans les résultats qui nous sont proposés est accrue du fait que nous les pensons personnalisés, et donc adaptés pour nous. Pariser affirme que les individus sont d'ores et déjà victimes de cette propagande des géants du web, et servent leurs intérêts économiques. Elle se manifesterait notamment par la mise en avant de façon dissimulée (et non pas dans des encarts publicitaires identifiés) de leurs propres produits ou de produits pour lesquels ils ont été payés. Antoinette Rouvroy et Thomas Berns appuient également cette idée, en soutenant que internet tend vers « *la captation systématique de toute parcelle d'attention humaine disponible au profit d'intérêts privés (l'économie de l'attention)* » [Rouvroy – Berns, 2013, p167]. On a en effet trop souvent tendance à penser que les entreprises utilisent le Big Data pour fournir une meilleure expérience aux utilisateurs et ainsi attirer une clientèle. Cela est en partie vrai, mais gagner une clientèle n'est intéressant pour eux que dans le cas où ils peuvent la monnayer, puisque les services qu'ils leurs offrent sont dans la plupart des cas gratuits. Cette monétisation s'effectue via la vente d'informations sur eux, ou de temps de cerveau disponible aux publicitaires.

Le Big Data ne serait pas utilisé seulement pour nous guider dans nos choix, mais de plus en plus pour les orienter. Selon Jessica Eynard, docteur en droit privé, **«** *la possibilité d'anticiper les comportements individuels grâce aux procédés techniques repose sur une logique simple : l'être humain étant un être rationnel, il est probable qu'il agisse de la même façon dès lors qu'il se trouve dans deux situations strictement identiques. Il suffit donc de l'observer pour pouvoir prévoir ses réactions. Or, plus on l'observe, plus on collecte d'informations sur lui et plus il devient facile de le rendre prévisible et d'anticiper, voire de susciter, ses réactions. Dans ce cadre, les informations personnelles deviennent la clé permettant d'orienter ses choix en fonction d'un profil établi informatiquement. »* [Eynard, 2012] Le profilage permettrait ainsi d'avoir une connaissance telle de l'individu qu'il permettrait de savoir comment susciter l'acte d'achat chez lui : savoir quel produit lui proposer, à quel moment et de quelle façon. Antoinette Rouvroy et Thomas Berns soulignent également les dérives inhérentes au profilage, qui selon ses défenseurs remettait le « *consommateur au centre des préoccupations des entreprises, [n'ayant] même plus à concevoir ni exprimer ses désirs qui sont des ordres.* » Selon eux, l'objectif consiste en réalité à « *adapter les désirs des individus à l'offre, en adaptant les stratégies de vente (la manière de présenter le produit, d'en fixer le prix…) au profil de chacun* », plutôt qu'à adapter l'offre en fonction des désirs de l'individu. L'exemple des sites de ventes de voyages aériens est selon eux très révélateur de cette tendance : elles cherchent à pousser l'utilisateur à l'achat, en adaptant leur offre en fonction du besoin de l'individu, évalué grâce à son mode de navigation. Si par exemple un individu

lambda cherche sur le site internet d'une compagnie un vol pour une échéance très courte, puis ferme la page et essaye de trouver ailleurs un meilleur prix avant de revenir sur ce premier site, la compagnie captera le fait que l'utilisateur est pressé puisqu'il cherche un vol d'urgence, et qu'il n'a pas trouvé mieux ailleurs puisqu'il est rapidement revenu sur ce site. Elle lui proposera alors le même vol majoré de quelques dizaines d'euros ; l'individu, déjà pressé, se sent d'autant plus pressé qu'il pense désormais que les prix augmentent vite du fait d'une forte demande, et va donc procéder à l'achat sans attendre. Les sites de vente de services ou produits utilisant le Big Data cherchent donc à créer un état d'alerte chez l'individu, afin de modeler son désir à leur offre et d'altérer son jugement.

Une autre critique couramment évoquée de la personnalisation sur internet souligne qu'elle n'est rendue possible essentiellement uniquement parce que les sites web captent de très grandes quantités d'informations sur leurs utilisateurs, et ce de façon souvent opaque et à leur insu, ce qui constitue une forte entrave au respect de la sphère privée. Le Big Data tire en effet ses informations des interactions entre les utilisateurs connectés et les objets qui les entourent (transactions en ligne, actions sur leurs *smartphones*, objets connectés, …), interactions qui constituent déjà en soi des informations d'ordre très privées. Mais là où le Big Data va encore plus loin, c'est qu'il combine et analyse, comme le soulignent Kate Crawford et Jason Schultz, ces sources hétérogènes d'information pour en déduire d'autres. Aussi, certaines données à priori inoffensives peuvent se révéler néfastes pour le respect de la vie privée lorsqu'elles sont combinées à d'autres. Les deux chercheurs nous donnent l'exemple de l'analyse des habitudes de consommation d'individus qui ont permis d'identifier parmi eux ceux qui attendaient un enfant. Des données à priori peu confidentielles, à savoir les achats effectués par un groupe d'individus, ont permis, après combinaison entre elles et passage par les algorithmes prédictifs, de déduire une information d'ordre intime, à savoir lesquels parmi eux attendaient un enfant.

Le problème avec les outils d'analyse du Big Data repose sur un paradoxe : les algorithmes prédictifs ne sont pas prévisibles, et il est donc extrêmement difficile de limiter leur portée et de légiférer dessus. Ce paradoxe constitue un obstacle pour le législateur, mais correspond également à un problème plus global de la nature humaine, explicité par Georg Simmel, philosophe et sociologue allemand, qu'il appelle la « *tragédie de la culture* »[7] : l'homme est poussé par nature vers la culture, il s'efforce donc de créer des choses qui lui sont dans un premier temps inutiles, puis qui lui deviennent indispensables et dont il devient dépendant. Il devient en quelque sorte l'esclave de sa création, il travaille pour elle, alors qu'il l'avait à l'origine conçue pour lui. On peut ici faire le parallèle avec les algorithmes prédictifs dont l'homme ne sait plus limiter la portée, mais ne veut plus s'en passer et continue de les alimenter ; il ne sont en quelque sorte plus à son service, mais il continue à les servir. Sous couvert de fournir une offre censée mieux répondre aux attentes de chacun des utilisateurs, les résultats et offres commerciales proposés sont déjà orientés, et cela implique des conséquences sur les annonces que chacun voit, les offres promotionnelles qui sont envoyées, les visions du monde qui sont proposées en termes de marketing et de publicité, … Mais cela commence aussi et surtout à influer sur les choix qui sont proposés et les opportunités

[7] *La tragédie de la culture* est un ouvrage de Georg Simmel, dans lequel il traite de cette question.

auxquelles chacun a accès, ce qui a des implications très importantes sur la vie des individus, et pose notamment des problèmes de discrimination. Ici, Crawford et Schultz nous donnent un autre exemple pour appuyer leur propos : dans le cas des locations immobilières, la pré-sélection des candidats est interdite et les législateurs sont là pour s'en assurer (on n'a pas le droit de spécifier sur les publicités qu'on n'accepte que les femmes par exemple), mais avec le Big Data ces contrôles sont extrêmement difficiles à mettre en place, car il permet, grâce à ses outils d'analyse, d'identifier les internautes qui correspondent au profil désiré (par exemple les individus du genre féminin), et par conséquent de les cibler dans la diffusion des annonces publicitaires. Ainsi, en fonction de son profil en ligne, un individu n'aura accès qu'à certaines offres commerciales, de location, d'emploi, … On constate ici une restriction du champ des possibles qui s'offrent à chacun, pour tendre vers un monde dans lequel on ne nous proposera que les opportunités dont on a estimé qu'elles correspondent, ou conviennent, à notre profil. Il nous semble ici important de remarquer que pourtant, à l'origine de la démocratisation de l'identité numérique, de nombreux discours la louaient en l'annonçant comme offrant la possibilité d'un nouveau départ pour les individus, permettant l'aplanissement des inégalités sociales. Cette identité peut sans doute s'expliquer par le fait qu'au départ, les profils étaient vierges et chaque individu était libre de l'agrémenter comme il le désirait. Nous remarquerons d'ailleurs que le même type de discours avait été énoncé avec l'arrivée d'internet, notamment justifié par le fait qu'internet favoriserait l'accès pour tous au savoir planétaire. Nous assistons à l'heure du Big Data, à un retour accentué, puisque plus difficilement contrôlable, de ces déterminations et de ces inégalités via l'univers numérique.

Les individus ne sont toutefois pas totalement passifs et ont également une influence sur les algorithmes du Big Data. Ces algorithmes reposent sur une acceptance du public quant à la pertinence des résultats qu'ils fournissent ; les services du Big Data sont constamment en train d'adapter leurs algorithmes en fonction des contenus qui sont appréciés ou pas par le public (« like », nombre de clics, taux de rebond, …). Les utilisateurs jouent ainsi un rôle important dans le sens où les résultats sont pertinents à partir du moment où ils les considèrent et les acceptent comme. Les acteurs du Big Data ont à tout prix besoin de conserver leur légitimité, et s'adaptent sans cesse aux exigences de leurs utilisateurs. On ne peut donc pas considérer les algorithmes comme une entité à part, mais plutôt la relation qui existe entre ces algorithmes et les utilisateurs, qui sont tous deux des variables évolutives, c'est pourquoi on parle d'influence mutuelle et récursive entre eux [Gillespie, 2012, p17]. L'exemple de Flickr, que nous rapporte Tarleton Gillespie, est également révélateur de ce phénomène d'influence mutuelle : Flickr propose des contenus différents en fonction des profils de leurs utilisateurs ; certains photographes s'en sont aperçus et s'adaptent donc aux spécificités de l'algorithme pour faire voir leurs photos par le public qu'ils visent (selon qu'ils cherchent un emploi, à vendre leurs photos, …). On a ici un exemple typique d'utilisation d'un algorithme par un utilisateur pour servir son propre dessein, mais force est de constater que l'algorithme oriente les choix esthétiques des photographes, et qu'il exerce par ce biais lui aussi une influence sur l'artiste. L'idée d'évaluation de la pertinence par l'utilisateur est toutefois nuancée par certains, qui affirment que les résultats sont la plupart du temps acceptés comme pertinents pour eux par les utilisateurs justement parce qu'ils leur sont servis avec l'idée implicite qu'ils

leur sont précisément et tout spécialement destinés. Le rapport de force pencherait donc très largement en faveur des entreprises qui fournissent les services.

L'ensemble des éléments que nous avons abordés dans cette partie nous ont permis de voir que le Big Data a bel et bien une influence sur nos comportements, nos modes de consommation, nos pratiques sociales, ainsi que la majorité des aspects de notre vie, dans le sens où ils ont changé notre façon de les aborder en nous obligeant à nous conformer à leurs modèles. Mais dans le même temps, nous avons pu mettre en lumière le fait que cette influence n'est pas à sens unique, et que l'individu possède lui aussi un certain pouvoir d'influence sur le Big Data. Le problème restant évidemment l'inégalité du rapport de force entre des individus isolés d'une part, et quelques grosses sociétés par lesquelles transitent quatre-vingt-quinze pourcents des pages consultées sur internet. Nous avons également vu les dérives auxquelles le Big Data peut mener, telles que la manipulation des individus et l'exploitation de leurs données personnelles.

Nous aborderons dans la partie suivante les changements que le Big Data a pu apporter dans la pratique scientifique et notamment dans le domaine de la recherche.

Science et recherche

Avec l'affluence des objets connectés, le Big Data collecte de plus en plus de données, notamment les informations relatives à la santé (mode de vie, activité physique, rythme cardiaque, tension, …), très sensibles et personnelles, mais qui peuvent néanmoins être potentiellement utilisées pour la recherche, particulièrement dans le domaine médical. L'analyse de grandes quantités d'informations sur les patients permettra aux algorithmes de faire des diagnostiques prédictifs et des suggestions de traitements préventifs, ce qui permettrait d'éviter la contraction de certaines maladies. Le *Research Kit* d'Apple, utilisé par les scientifiques pour créer des applications sur-mesure en fonction de leurs recherches afin de recueillir très rapidement des informations sur un très grand nombre d'utilisateurs (potentiellement à terme tous les utilisateurs d'iPhone), est révélateur de cette tendance.

Alors qu'autrefois le recueil d'informations par les chercheurs sur quelques milliers de personnes était un processus très long à mettre en place, le Big Data permet aujourd'hui de recueillir des données provenant potentiellement de toute une population, et ce de façon presque instantanée. Notre capacité à stocker des quantités infinies de données et à les comprendre serait donc en train de révolutionner la science et la médecine. A titre d'exemple, l'analyse par des algorithmes prédictifs d'un ensemble complet et hétérogène de données sur un patient (liste des traitements pris au cours de la vie et qualité de la réponse à ceux-ci, échantillons biologiques, antécédents familiaux, génome de la flore intestinale, indice de masse corporelle …) croisées avec nos connaissances médicales permettraient, à terme, d'estimer avec une grande précision le risque de subir un arrêt cardiaque, un cancer, ou toute autre maladie, ainsi que de proposer des solutions pour éviter ou limiter les risques de contracter ces maladies (changement de régime alimentaire, traitement préventif, …). L'application du Big Data à la médecine est d'ailleurs déjà effective puisque le super-ordinateur Watson, fabriqué par IBM, est utilisé depuis 2012 dans le traitement du cancer du poumon au *Memorial Sloan-Kettering Cancer Center* à New York. L'ordinateur, capable

grâce aux méthodes sémantiques du Big Data, de comprendre le langage naturel et d'interpréter et d'analyser des informations très variées, compare les données biologiques et corporelles des patients avec un très grand nombre d'articles scientifiques, afin de proposer aux médecins des hypothèses de traitements les plus adaptés aux spécificités de chacun d'entre eux. De même, Google cherche depuis quelques années à collecter les séquences ADN des individus afin de les entrecroiser et de les corréler avec les dossiers médicaux de chacun d'entre eux pour deviner statistiquement les maladies qu'ils sont susceptibles de développer.

Mais par-delà même les possibilités nouvelles offertes par le Big Data à la science, c'est la méthode scientifique elle-même qui serait remise en question. La méthode scientifique traditionnelle repose en effet sur un processus rigoureux, au cours duquel le chercheur émet des hypothèses théoriques qu'il cherche à confirmer ou à infirmer en suivant un modèle. Le modèle définit les moyens expérimentaux permettant d'éprouver la valeur des hypothèses, à savoir les expérimentations qui seront mises en œuvre et les données qu'il s'agira de collecter et d'exploiter. L'application des algorithmes prédictifs, qui présuppose que l'opportunité de trouver des réponses à des questions fondamentales augmente à mesure que leur collection de faits croît. Mais à très grande échelle, les données ne peuvent plus être visualisées dans leur totalité, et on doit utiliser les algorithmes pour les analyser, les associer et les utiliser dans un contexte. La méthodologie scientifique traditionnelle est donc ici bouleversée, dans le sens où on n'analyse plus le sens des données, mais les résultats statistiques fournis par les algorithmes. De plus, autre bouleversement, on peut désormais analyser des données sans hypothèses sur ce qu'elles pourraient révéler, et laisser les algorithmes nous révéler des choses inattendues, sur lesquelles aucunes suppositions n'ont été préalablement établies. Le processus est en fait ici inversé, puisque ce sont les données elles-mêmes qui permettent la formulation d'hypothèses, contrairement au modèle traditionnel, qui utilise des informations dans le but de vérifier une hypothèse préalablement définie. Ce phénomène est d'ailleurs très justement souligné par Antoinette Rouvroy et Thomas Berns : « *on constate que les « savoirs » produits ont comme principale caractéristique de paraître émerger directement de la masse des données, sans que l'hypothèse menant à ces savoirs ne leur préexiste : les hypothèses sont elles-mêmes « générées » à partir des données* » [Rouvroy – Berns, p180, 2013].

Pour Chris Anderson, administrateur de la conférence TED[8], ce bouleversement des méthodes scientifiques ne pose pas de problème, et constitue au contraire une avancée : « *Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.* »[9] [Anderson, p1] Selon lui, les modèles constituaient plus une limite à la science qu'un vecteur de progression, et n'étaient utilisés que par faute d'autres alternatives. Aujourd'hui, les technologies du Big Data permettent de s'en affranchir, et donc de dépasser les imperfections liées aux modèles. Anderson nous dit que l'approche traditionnelle selon laquelle les données sans modèle ne constituent en rien une preuve est désormais dépassée avec l'arrivée du Big

[8] *« Les conférences TED (Technology, Entertainment and Design), sont une série internationale de conférences organisées par la fondation à but non lucratif Sapling foundation. Cette fondation a été créée pour diffuser des « idées qui valent la peine d'être diffusées » »* - Wikipédia

[9] Les sociétés actuelles comme Google, qui ont grandi dans une période d'abondance massive des données, n'ont pas à se limiter à de mauvais modèles. Ils n'ont d'ailleurs pas à s'encombrer de modèles du tout

Data. Selon lui, les modèles scientifiques (lois de Newton en physique, théories sur les gènes dominants et récessifs en biologie, …), ont tous montré leurs limites et on sait aujourd'hui qu'ils ne constituent qu'une approximation d'une réalité plus complexe. Autrement dit, le plus on progresse en science et le plus on s'éloigne d'un modèle unique qui peut l'expliquer [Anderson, 2008], phénomène qui peut être expliqué par le fait que plus on accumule des données, plus il est complexe de les faire rentrer dans un modèle unique à portée générale, la diversité de la réalité empirique remettant toujours en cause les cadres théoriques préétablis. Alors que le Big Data, plus pragmatique, ne cherche pas à nous expliquer pourquoi telle ou telle chose se produit, mais énonce des faits appuyés par des statistiques réalisées sur une gigantesque masse de données, dont la fiabilité est très forte ; lorsque la quantité de données est suffisante, les chiffres parleraient donc d'eux-mêmes. Pour appuyer son propos, Anderson prend l'exemple de John Craig Venter, biologiste et directeur du centre pour la recherche en génomique[10], qui à l'aide de superordinateurs, a fini par séquencer[11] génétiquement la plupart de l'océan, découvrant ainsi des milliers d'espèces de bactéries auparavant inconnues. Ses recherches ne permettent pas d'expliquer l'origine de ces espèces, de quel ancêtre elles descendent ou quelle était leur utilité dans l'équilibre naturel à ce moment-là, comme auraient cherché à l'identifier les chercheurs via les modèles traditionnels, mais il n'en reste pas moins qu'il a fait avancer la science simplement en utilisant les méthodes du Big Data.

S'il est en effet bénéfique pour l'avancement de la recherche, notamment médicale, d'avoir accès à moindre coût, facilement et dans des proportions jamais connues jusqu'alors aux données individuelles, le fait que les entités qui collectent et sont capables d'exploiter et d'analyser ces données soient essentiellement des sociétés privées pose d'importants problèmes en termes de protection de la vie privée. On constate que les états eux-mêmes font appel à des prestataires privés pour collecter certaines données relatives à la santé, et ce sans demander aucune garantie. Il leur serait d'ailleurs extrêmement difficile d'obtenir des garanties suffisantes et fiables, car les données sont stockées de façon totalement opaque[12], à moins d'exiger justement une transparence quant à la façon dont les données sont mémorisées et exploitées ; ces prestataires peuvent donc, une fois en possession des données, les exploiter à titre privé ou les revendre. Les données relatives à la santé notamment, sont pourtant très sensibles car elles relèvent de ce qu'il y a de plus intime pour l'individu, et nécessitent d'être protégées. Mais dès lors qu'elles sont collectées et exploitées de façon complètement opaque par Apple par exemple, aucun moyen ne s'offre à l'individu pour s'assurer que l'usage qui en est fait n'a pas un but commercial ou détourné par rapport à celui pour lequel elles ont été fournies. Selon la législation française, les données médicales sont la propriété de celui dont elles émanent, mais garantir ce droit de propriété à l'individu est devenu très complexe. Afin de préserver la confidentialité de ces données et prévenir les risques et les déviances qui peuvent survenir, il est donc essentiel d'établir des processus pour identifier et contrôler ces données, et encadrer l'usage qui en est fait. On imagine en effet aisément l'implication que pourrait avoir le fait que des sociétés comme les mutuelles puissent avoir accès au profil

[10] The Center for the Advancement of Genomics (TCAG)

[11] « Détermination de la séquence (suite ordonnée) des bases azotées constituant un fragment d'A.D.N » - Larousse. Et dans le cas présent le séquençage s'étend au génome complet.

[12] Voir par exemple le récent scandale autour des appareils à pression positive continue pour prévenir l'apnée du sommeil, ou encore le cas de l'Islande qui a vendu les données génétiques de sa population à une société privée.

médical détaillé d'un individu, combiné aux prédictions offertes par le Big Data permettant d'identifier les risques pathologiques qu'il présente. Les individus identifiés comme « à risque » pourraient ainsi à l'avenir se voir refuser une couverture par les mutuelles, car leurs profils ne convergeraient pas avec leurs intérêts économiques.

Au-delà des questions éthiques que pose évidemment cette volonté de transformer internet, et la quantité gigantesque de données potentiellement sensibles et à caractère privé qu'il contient, en « *laboratoire de notre condition humaine* » [Anderson, 2008], certains chercheurs comme George Dyson, historien des sciences, émettent des réserves quant au fait que la méthode traditionnelle telle qu'on la connaît puisse devenir obsolète et disparaître. Le Big Data est en effet déjà utilisé dans de nombreux domaines (comme l'astronomie ou la physique), sans avoir pour autant remplacé les anciennes méthodes, et vient plutôt les complémenter. La collecte et l'analyse de données ont toujours fait partie de la méthode scientifique, on ne peut par conséquent pas parler, selon William Daniel Hillis, ingénieur et mathématicien américain, de révolution, alors qu'il s'agit d'un simple changement d'échelle. La démultiplication des capacités de collecte et d'analyse est donc un atout de taille, mais seulement si elle est utilisée à bon escient et par des scientifiques compétents, qui l'orientent dans le sens de leurs recherches et pour augmenter leur propre potentiel d'analyse et de création. Néanmoins les machines ne vont pas remplacer le raisonnement et le travail humains, l'humain étant seul capable de se projeter et d'orienter ses recherches. Antoinette Rouvroy et Thomas Berns doutent également de la possibilité et du bien-fondé de la pratique de la science sans émission d'hypothèses. Cette absence ferait en effet disparaître l'idée de « *ratés* », qui permettent à un projet d'être « *repris et transformé* », d'évoluer. Cette idée a d'ailleurs été énoncée dès le XXème siècle, avec la théorie du *falsificationnisme* du philosophe des sciences Karl Popper : selon lui, on doit chercher l'invalidation d'une théorie de manière expérimentale, en trouvant un cas d'étude qui va à son encontre. La science progresserait en falsifiant d'anciennes théories au bénéfice de nouvelles, elle serait donc basée sur des erreurs, des « *ratés* », ouvrant la voie à de nouvelles théories. Dans le cas de la recherche scientifique telle que la prône Anderson, le chercheur ne fait qu'attendre que les algorithmes prédictifs lui révèlent des avancées, il ne formule donc plus d'hypothèses, alors que ce sont justement les hypothèses qui sont mises à l'épreuve par la méthode scientifique et peuvent se révéler fausses ; sans hypothèse l'erreur, le raté disparaissent. Le raté est pourtant à la base de la vie organique, et ainsi de l'ensemble des choses qui nous entourent, y compris du processus scientifique, qui cherche et doit se calquer sur le vivant. Ce sont en effet les hommes, qui sont avant tout des êtres vivants, qui font la science. Il est donc normal d'y retrouver cette dimension d'imprévu, de non exhaustivité, d'instabilité, d'adaptation et d'évolution dans les théories. : « *or, même pour un organisme, même pour la vie, pour l'organique en tant que lieu d'une activité normative, il y a du raté, du conflit, du monstrueux, de la limite et du dépassement de la limite, avec les déviations et les déplacements que cela induit dans la vie* ». [Rouvroy – Berns, p182, 2013]

Par ailleurs, John Horgan, journaliste scientifique américain, nous fait remarquer que la théorie de Chris Anderson est limitée dans le sens où elle se contente de réduire la science à la prédiction de l'avenir grâce au passé. Les algorithmes du Big Data ne reposent en effet que

sur des événements et des informations qui ont été enregistrés, et qui par conséquent appartiennent au passé. Seul l'humain est capable d'anticiper et de réagir à des phénomènes nouveaux et inattendus. L'humain garderait donc toujours une longueur d'avance sur les algorithmes, à savoir sa capacité à se projeter, à imaginer l'avenir à partir d'idées nouvelles et non d'événements antérieurs.

Enfin, il nous semblait important d'aborder l'idée de Chris Anderson, selon laquelle la recherche scientifique pratiquée par les algorithmes est plus performante parce qu'elle annihile les biais liés à l'interprétation et aux modèles construits par les humains. Mais pour autant, la vision de la science proposée par Anderson ne se soustrait pas à toute trace de subjectivité ou de biais, car les algorithmes sur lesquels elle repose, comme nous le verrons dans la prochaine partie, sont révélateurs d'une certaine vision du monde et ont été construits dans un but et une orientation précis par ceux qui les ont développés. Le biais introduit est donc moins direct mais existe toujours bel et bien. Dans le cas de la recherche médicale s'appuyant sur les informations transmises par les utilisateurs via des objets connectés, se pose également la question de savoir si les informations collectées ne sont pas biaisées car elles proviennent uniquement d'individus disposant de *smartphones*, qui appartiennent peut-être uniquement à une certaine catégorie, aisée, favorable aux innovations technologiques, … On peut certes imaginer que ce biais va potentiellement tendre à se réduire du fait que de plus en plus de gens posséderont des *smartphones* à l'avenir, mais il n'est pas négligeable à l'heure actuelle.

La métaphore d'Anderson, parlant d'internet comme le « *laboratoire de notre condition humaine* », implique d'autres conséquences que celles uniquement liées à la recherche et à la science. En effet, si internet est le reflet des comportements et des instincts humains, il peut devenir le lieu d'enjeux cruciaux pour les gouvernements, dans leur objectif de gérer et catalyser les populations, car il leur permettrait d'anticiper et prévenir les comportements dits « déviants » ; nous aborderons donc dans la prochaine partie l'évolution des pratiques gouvernementales en lien avec le Big Data.

Evolution des pratiques gouvernementales

Nous utiliserons, dans cette partie, de manière récurrente le terme de « gouvernementalité ». Nous allons donc dans un premier temps définir ce que nous entendrons par ce concept. Nous parlerons de « gouvernementalité » en tant qu' « art du gouvernement » [Foucault, 1978]. Elle est une forme d'exercice du pouvoir, dont Michel Foucault, philosophe français qui a essentiellement travaillé sur les rapports entre pouvoir et savoir, identifie deux variantes. La frontière entre les deux est souvent floue, comme nous tâcherons de le montrer dans cette partie : la première est la pratique de la souveraineté, dont la finalité est d'obtenir des individus sous son autorité qu'ils « obéissent tous et sans défaillance aux lois » [Foucault, 1978], donc une soumission absolue de leur part ; dans ce cas l'objectif visé est uniquement le maintien de sa souveraineté par le souverain, et les moyens utilisés pour y parvenir sont essentiellement les lois et la répression. L'autre forme d'exercice du pouvoir correspond à ce que nous entendons par « gouvernementalité », et sa finalité consiste à conduire à une « *fin convenable* » par la « *disposition des choses* » [Foucault, 2004]. Nous reprendrons la définition de Michel Foucault, qui utilise judicieusement l'image de la famille, et pense

qu'elle consiste, historiquement, à « *répondre essentiellement à cette question : comment introduire l'économie, c'est-à-dire la manière de gérer comme il faut les individus, les biens, les richesses comme on peut le faire à l'intérieur d'une famille, comme peut le faire un bon père de famille qui sait diriger sa femme, ses enfants, ses domestiques, qui sait faire prospérer la fortune de sa famille, qui sait ménager pour elle les alliances qui conviennent, comment introduire cette attention, cette méticulosité, ce type de rapport du père de famille à sa famille à l'intérieur de la gestion d'un État ?* » [Foucault, pp12-29]. La gouvernementalité consisterait donc à gérer de façon optimale l'ensemble des éléments qui se trouvent sur le territoire gouverné, à savoir évidemment les individus qui y vivent et les richesses qui s'y trouvent, mais aussi tous les facteurs extérieurs, comme les catastrophes naturelles, les relations diplomatiques avec les gouvernements voisins, … La notion de gouvernementalité, contrairement à celle de la souveraineté, introduit par ailleurs les notions de bien-être et de bien commun dans l'exercice du pouvoir. Nous remarquerons également que les moyens pour exercer ce pouvoir sont fondamentalement différents de ceux de la souveraineté qui résident essentiellement dans la l'exercice de la loi du plus fort. Dans le cas de la gouvernementalité, le gouvernant doit chercher une « *disposition [judicieuse] des choses* » pour parvenir à un équilibre, il doit donc faire usage de tactiques pour amener les individus à se comporter de la façon la plus profitable pour tous, et non faire usage de la loi.

Le Big Data a profondément transformé les stratégies et les tactiques de gouvernement, notamment avec ce qu'on appellera la « *gouvernementalité algorithmique* » [Rouvroy – Berns, 2013], qui repose sur le fait que dans les sociétés contemporaines, de très grandes quantités de données sont collectées, corrélées et analysées automatiquement, dans le but d'anticiper les comportements des citoyens. Il s'agira donc de montrer dans cette partie ce que permettent les nouvelles pratiques gouvernementales rendues possibles par le Big Data, et si elle mènent à des dérives, dans le sens où elles s'approcheraient plus de la pratique de la souveraineté que de celle de la gouvernementalité.

Aux Etats-Unis, les agences chargées du maintien de l'ordre commencent à se tourner vers les méthodes prédictives du Big Data : elles utilisent la date, l'heure et la localisation des crimes récents, qu'elles combinent avec les archives des crimes, dans le but d'identifier des « *zones à* risque » [Crawford - Schultz, 2014, p103] sur lesquelles vont se concentrer les forces de l'ordre (vidéo-surveillance, patrouilles, …). On cherche dans un premier temps à identifier un certain nombre de critères permettant d'évaluer la dangerosité d'un endroit, suite à quoi on tente de prédire la probabilité qu'un crime soit observé à un endroit et à un moment donnés, afin de répartir les patrouilles de police en fonction des résultats obtenus. Les agences espèrent prévenir les crimes à venir grâce à cette présence policière plus importante dans les zones ciblées, et même résoudre des affaires passées car les méthodes du Big Data permettent facilement de relier des individus à des informations sur le lieu et la date, recoupées avec les informations judiciaires disponibles (casiers judiciaires, …). Les individus qui vivent dans ces zones à risque seront donc sujets à une surveillance accrue de la part de la police, et potentiellement considérés comme suspects du fait de leur lieu de vie, ce qui constitue déjà une atteinte à leur vie privée, et à la présomption d'innocence. Mais le principal problème, selon Kate Crawford et Jason Schultz, est que dans les zones dites « *à risque* », moins

d'infractions passeront inaperçues, contrairement aux zones identifiées comme sûres, à cause de la surveillance renforcée mise en place. L'augmentation du nombre d'arrestations et d'infractions constatées alimentera donc les archives de la police pour ces zones, qui se trouveront enfermées dans leur statut de « *zones à risque* », car les algorithmes de prédiction s'appuient sur ces archives. Les individus qui y vivent se trouveront donc prisonniers de cette situation de surveillance accrue, ce qui pose bien évidemment un problème de discrimination. On fait face ici au même type de problématique que lors du débat houleux qui avait eu lieu il y a quelques années à propos du « délit de faciès ». Le délit dans le cas présent n'étant pas l'apparence physique mais le lieu de vie.

Les agences de surveillance et les gouvernements auxquels elles sont rattachées assurent pour la plupart d'entre elles, et notamment dans les pays démocratiques, qu'elles préservent l'anonymat des personnes surveillées et n'alimentent leurs algorithmes que des données qu'ils laissent sur leur passage. Toutefois, cette garantie s'avère peut convaincante car il est très souvent possible de déduire l'identité d'une personne à partir des traces qu'elle a laissées. Nous noterons d'ailleurs que la réciproque est vrai, et qu'il est possible de profiler une personne avec une très forte précision à partir de ses traces de navigation. Peut-être même avec plus de précision qu'en analysant le contenu des données qu'il transmet, car les traces, à l'inverse des données publiées, ne mentent pas.

Les agences de surveillance américaines (CIA, FBI et armée notamment) font d'ailleurs quant à elles une utilisation du Big Data beaucoup plus poussée, en faisant appel à des centres de partage de l'information appelés « *fusion centers* », afin de regrouper et agréger leurs données respectives. Ces données sont ensuite analysées afin d'identifier des individus suspects ou qui méritent d'être surveillés. Or il se trouve que ces méthodes peuvent mener à des erreurs grotesques. L'exemple choisi par les deux chercheurs pour illustrer ces erreurs est d'ailleurs frappant :

« In one case, Maryland state police exploited their access to fusion centers to conduct surveillance of human rights groups, peace activists, and death penalty opponents over a nineteen-month period. Fifty- three political activists eventually were classified as "terrorists", including two Catholic nuns and a Democratic candidate for local office. The fusion center shared these erroneous terrorist classifications with federal drug enforcement, law enforcement databases, and the National Security Administration, all without affording the innocent targets any opportunity to know, much less correct, the record. » [13] [Crawford – Schultz, 2014, p104]

Cet exemple met en lumière le fait que les agences de surveillance ciblent des groupes d'individus sans raison valable apparente, et d'ailleurs visiblement sans avoir à se justifier auprès d'aucune instance avant de mettre en place cette surveillance, puis exploitent et

[13] « *La police d'état du Maryland a utilisé ces centres pour surveiller des associations pour la défense des droits de l'homme, des activistes pour la paix, et des opposants à la peine de mort pendant dix-neuf mois. Cinquante-trois activistes politiques ont été identifiés comme « terroristes », dont deux nonnes catholiques et un candidat démocrate pour un poste local. Ces résultats erronés ont ensuite été partagés avec l'agence de lutte contre le trafic et drogue et l'agence nationale de sécurité, sans même offrir aux innocents visés l'opportunité de prendre connaissance, et encore moins de corriger, le dossier transmis.* »

partagent les résultats obtenus sans même en informer les personnes visées, alors qu'elles encourent des poursuites judiciaires. Avec ces méthodes de prédiction, le gouvernement collecte non seulement de grandes quantités d'informations sur les individus, mais génère grâce à ses outils d'analyse des données qui mettent à jour des informations intimes sur les individus. Crawford et Schultz soulignent d'ailleurs que même les sociétés qui développent les logiciels de prédiction utilisés par les agences de sécurité admettent qu'ils ne respectent pas les règles éthiques de respect de la vie privée. On voit donc ici que la collecte massive de données et son exploitation par des organismes gouvernementaux, en plus de ne pas respecter l'intimité des individus, peuvent avoir de très graves conséquences pour eux, prouver son innocence après avoir été identifié comme terroriste peut en effet s'avérer très complexe, surtout aux Etats-Unis et dans le contexte actuel.

Ces erreurs peuvent trouver leur explication dans ce que Antoinette Rouvroy et Thomas Berns appellent les « *faux-positifs* » [Rouvroy – Berns, 2013], et constituraient l'essence même du fonctionnement des algorithmes prédictifs. Selon eux, la gouvernementalité algorithmique tire sa force du fait que « *plus on s'en sert, plus le système algorithmique s'affine et se perfectionne, puisque toute interaction entre le système et le monde se traduit par un enregistrement de données numérisées, un enrichissement corrélatif de la « base statistique », et une amélioration des performances des algorithmes* » [Rouvroy – Berns, 2013, p174]. Cela entraîne le fait que le « *raté* » déjà évoqué pour notre partie sur la science, n'est pas considéré comme tel par les algorithmes, puisque toute interaction avec le système, même défectueuse, permet de l'enrichir. Cela pose un problème au niveau des pratiques gouvernementales, car concernant la sécurité, les « *faux-positifs* », à savoir les individus identifiés comme criminels ou terroristes par les algorithmes alors qu'ils sont en réalité innocents, ne seront pas considérés comme des erreurs, l'objectif des algorithmes étant « *de ne rater aucun vrai positif, quel que soit le taux de faux-positifs* » [Rouvroy – Berns, 2013, p174]. La gouvernementalité algorithmique, en cherchant à prévenir tous les risques, s'autorise donc l'erreur judiciaire. Antoinette Rouvroy et Thomas Berns ne nient pas les aspects positifs que peut apporter le Big Data en matière de gouvernementalité et notamment de sécurité, mais ils nous mettent en garde, et rappellent le devoir des gouvernants « *d'éviter que des décisions produisant des effets juridiques à l'égard de personnes ou les affectant de manière significative ne soient prises sur le seul fondement d'un traitement de données automatisé* » [Rouvroy – Berns, 2010, p170]. Les « *savoirs* » apportés par les algorithmes ne sont en effet constitués « *que de corrélations* », qu'il faut considérer et ne pas dénigrer, mais qui nécessitent une certaine réserve à leur égard : il faut garder à l'esprit qu'une corrélation ne constitue pas une relation de cause à effet, comme on a trop tendance à le penser.

Le refus de l'erreur, du « *raté* » pose également une question d'ordre philosophique, car ce postulat annihile toute mise à l'épreuve des individus, mais aussi et surtout des normes et des règles. Le gouvernement algorithmique cherche à prévoir et anticiper tous les comportements afin de prévenir ceux qui sont identifiés comme indésirables, d'empêcher qu'ils ne se produisent en les rendant « *physiquement impossibles* » [Rouvroy – Berns, 2010, p101]. Le mode d'exercice du pouvoir est ainsi totalement bouleversé puisqu'on ne cherche plus à exercer des « *contraintes* » sur les individus afin de les empêcher ou les dissuader d'agir de

façon indésirable, mais plutôt à modifier l'environnement afin de rendre ces comportements indésirables impossibles. Le gouvernement statistique, contrairement aux formes de gouvernement classiques qui n'agissent que sur les infractions passées, se concentre donc sur les événements à venir. Cet état de fait peut apparaître à priori comme étant positif, et l'est d'ailleurs du point de vue de la loi et de la sécurité, mais c'est omettre que la désobéissance constitue la résistance aux lois, et permet leur remise en question. En effet, s'il n'est plus possible de ne pas respecter les règles, les lois ne sont plus jamais outrepassées, et ainsi plus remises en question puisqu'on ne les sollicite plus ; nous remarquerons ici que ce type de régime s'apparente fortement à un régime totalitaire, dans le sens où il annihile toute forme d'opposition à la loi. L'absence de désobéissance pose également un autre problème majeur, celui du « *déséquilibre [de] l'architecture démocratique* » [Rouvroy – Berns, 2010, p102]; en effet si plus aucun acte illégal ne peut être commis, alors les instances juridictionnelles n'ont plus lieu d'être. Or celles-ci sont selon nous constitutives de la démocratie, fondée sur le principe de séparation des pouvoirs, notamment de la séparation des pouvoirs législatif et judiciaire.

Nous avons donc montré dans cette partie que la gouvernementalité algorithmique pouvait apporter un certain nombre de bienfaits, notamment en termes de sécurité, mais que certaines barrières ne devaient pas être franchies au risque d'attenter non seulement aux libertés individuelles, mais aussi et surtout à la démocratie elle-même, et donc à la légitimité du gouvernement. De la même façon que dans le cas de la science abordé dans la partie précédent, la thèse de la pureté théorique de la seule récolte massive de données peut être remise en question du fait qu'il y a toujours un ciblage qui s'opère et détermine les données qui auront une valeur théorique, ce qui implique des conséquences dans les résultats fournis par les algorithmes. Cette notion de biais des algorithmes sera d'ailleurs principalement l'objet de la partie qui suit. Nous conclurons, de la même façon que nous avons commencé, avec une citation de Michel Foucault, qui estime qu'il doit exister dans un état une « *continuité ascendante, en ce sens que celui qui veut pouvoir gouverner l'État doit d'abord savoir se gouverner lui-même ; puis, à un autre niveau, gouverner sa famille, son bien, son domaine, et, finalement, il arrivera à gouverner l'État* » [Foucault, 1978, pp12-29]. Pour pouvoir exploiter à bon escient et légiférer comme il se doit sur le Big Data, le gouvernant doit donc avant tout être à même de l'appréhender et connaître les enjeux qu'il implique tout d'abord pour lui en tant qu'individu, mais aussi pour ses proches, et enfin de les considérer pour l'ensemble des individus qu'il aspire à gouverner.

Nous avons abordé dans cette première partie les différents secteurs dans lesquels s'illustre particulièrement le Big Data, et montré ainsi qu'il intervenait dans la plupart des aspects de notre vie, autant publique que privée. Nous tâcherons donc, dans la partie suivante, d'expliquer comment cet état de fait a été rendu possible, et s'il est discutable.

Légitimité des algorithmes du Big Data

L'omniprésence du Big Data dans la vie privée des individus n'est rendue possible que par une certaine légitimité qui lui est conférée par le public, et une certaine acceptance tacite de cette omniprésence de la part des individus. Cette acceptance est en partie due au fait que les

individus considèrent que le Big Data présente pour l'instant plus d'avantages que d'inconvénients. En effet, en contrepartie de cette collecte des données privées, les individus estiment gagner un temps et un confort précieux au cours de leur navigation sur internet : ils ont accès à des services qui leur sont personnellement dédiés et s'adaptent à eux. Néanmoins cette seule explication ne suffit pas, c'est pourquoi nous allons tâcher dans cette partie de montrer sur quoi repose cette légitimité, et d'identifier ses limites.

Facteurs de légitimité

La première variable que nous aborderons, relevée par Antoinette Rouvroy et Thomas Bern, est la suppression de la signification de la donnée et de ce qui la rattache à son émetteur. Elle constitue une condition essentielle pour que le public comprenne et accepte que le Big Data capte de très grandes quantités de données personnelles.: Toutes nos données de navigation sont en effet, comme nous l'avons montré dans la première partie, collectées et conservées « par défaut », et ce la plupart du temps sans que l'individu s'imagine qu'il est en train de fournir de l'information. Par exemple, le seul fait d'ouvrir une page web constitue des traces qui sont stockées, sans pour autant qu'un internaute non averti puisse l'imaginer ; il s'agit donc plus de *« traces laissées »* que de « *données transmises* » volontairement [Rouvroy – Berns, 2013] ; pour autant, dans l'esprit de la majorité du public elles n'apparaissent pas comme « *volées* » [Rouvroy – Berns, 2013]. Cela s'explique par le fait que ces données sont complètement déconnectées du cadre dans lequel elles ont été collectées, puis corrélées à d'autres sans qu'on puisse prédire quelles informations elles fourniront et à quoi elles serviront. Les traces stockées sont détachées de l'utilisateur qui les a laissées, et sont stockées sans intention préalable, minorant ainsi le sentiment d'intrusion dans la vie personnelle de l'individu. Ces données stockées leur paraissent ainsi peu intrusives. Il est également intéressant de noter que cette séparation des données du contexte dans lequel elles ont été prélevées contribue également à l'impression d'objectivité des données et des algorithmes qui les exploitent, qui est souvent mise en avant par les acteurs du Big Data, et que nous allons développer ensuite.

Le principal motif invoqué par les acteurs du Big Data pour légitimer leurs pratiques est l'objectivité de leurs algorithmes. L'objectivité consisterait à s'abstraire du biais inhérent à l'intervention de l'homme dans la prise de décision. L'homme, consciemment ou malgré lui, est en effet toujours sujet à des influences extérieures (d'ordre économique, éthique, politique, …). L'objectivité, telle qu'entendue par les acteurs du Big Data qui s'en revendiquent, consisterait donc à l'absence de déterminations sensibles dans le choix d'une fin. Cette objectivité constitue, pour les individus, un gage de fiabilité et de justesse, qui permet d'éviter les erreurs ou les influences extérieures ; notamment économiques avec la mise en avant d'offres économiques qu'on nous présenterait comme des résultats répondant au mieux à notre requête. Si les algorithmes sont objectifs, alors les résultats qu'ils nous proposent sont effectivement les mieux adaptés et les plus pertinents en fonction de la connaissance qu'ils ont de nous. Obtenir le meilleur résultat qui soi pour lui, en éliminant toute trace de biais inhérent à l'homme, constitue souvent aux yeux de l'individu une compensation suffisante à la collecte de ses données personnelles. Mais il existe également une autre raison, moins mise en avant, à cette revendication : une telle objectivité des « machines » qu'ils ont créées leur

permet de se déresponsabiliser. Ces entreprises, et notamment Google, sont devenues les plus importants portails d'accès à l'information jamais connus. Comme nous le fait remarquer James Grimmelmann à très juste titre, les fournisseurs des principaux moteurs de recherche ont une « *influence énorme sur nous tous* » [Grimmelmann, 2008] : ils peuvent déterminer et influencer « *ce que nous lisons, qui nous écoutons, et qui est entendu* » [Grimmelmann, 2008]. Selon Evgeny Morozov, chercheur spécialiste des implications politiques et sociales du progrès technologique, cette responsabilité les embarrasse et elles ne veulent pas avoir à assumer les biais inhérents à leurs systèmes, qui peuvent avoir de graves conséquences sur la vie des individus et sur l'évolution de nos sociétés.

Afin d'illustrer cette volonté de la part des sociétés du Big Data de paraitre objectives, nous prendrons l'exemple de Google, très révélateur de cette tendance. Les premiers moteurs de recherche analysaient le contenu des pages web et mesuraient en leur sein le nombre d'occurrences du terme recherché. Ces moteurs pouvaient évidemment être trompés extrêmement facilement en écrivant un grand nombre de fois sur une page web les termes les plus couramment recherchés. Les fondateurs de Google, Larry Page et Sergey Brin, ont réussi à dépasser cette limite inhérente à la recherche lexicale, en s'intéressant non plus au contenu en tant que tel pour juger de la pertinence d'un site internet, mais aux « attributions respectives que les sites se font les uns envers les autres en se reconnaissant ». A savoir que plus le nombre de publications web faisant référence à une autre via un système de références appelées « hyperliens »[14], plus le contenu de ce site internet sera jugé pertinent, et donc mis en avant dans les résultats de recherche ; ce mode d'évaluation de la pertinence d'une publication via un système de notation sur une échelle de 1 à 10 est appelé le PageRank[15]. On estime par-là que plus un site web est cité par d'autres, plus ceux-ci l'ont jugé intéressant, et donc plus il mérite d'être visible. En effet, lorsqu'en tant qu'internaute, un individu publie un lien vers une publication, c'est qu'il la juge intéressante et souhaite attirer l'attention des autres internautes sur elle. Lorsqu'un site en cite un autre, il lui confère ainsi une certaine autorité. Le seul dénombrement de ces « transferts d'autorité » permettrait donc de mesurer l'autorité d'un site internet. Google utilise aussi ces liens pour déterminer de quoi traitent les pages web : en effet, comme nous le précise James Grimmelmann, professeur de droit à l'Université du Maryland[16], plus le nombre de publications faisant référence à une page web en utilisant une certaine phrase ou certains mots clefs sera important, et plus Google identifiera cette phrase ou ces mots-clefs comme décrivant précisément le contenu de cette page.
Le PageRank a permis à Google de revendiquer son algorithme comme « *un champion de la démocratie* » [Cardon, 20, p72]. Une citation peut être vue comme un vote auquel chacun peut participer (une publication permettant de créer un vote via l'intégration d'hyperliens en son sein). Cependant le PageRank de Google ne s'arrête pas là, et considère également l'autorité du citeur. Si celui-ci a une forte autorité (donc s'il est lui-même cité de nombreuses fois), alors sa citation aura plus de valeur qu'une citation émanant d'un site qui n'est jamais cité par aucun autre. Sur le web, chaque internaute qui publie devient un citeur et par là un

[14] Un hyperlien, ou lien web, permet « *le passage d'une page web à une autre à l'aide d'un clic* » - Wikipédia
[15] Pour une description détaillée du mode de fonctionnement du PageRank, voir l'ouvrage, cité en Bibliographie, de Dominique Cardon : « Dans l'esprit du PageRank : une enquête sur l'algorithme de Google »
[16] University of Maryland Francis King Carey School of Law

votant, cependant leurs publications ne sont pas d'égale qualité. Google cherche donc les citeurs qui ont été identifiés par d'autres comme de « bons » citeurs (qui ont été cités de nombreuses fois), afin de leur donner un poids plus important. Ce dernier point permet de s'assurer que les publications mises en avant dans les résultats sont de qualité, en évitant notamment que les internautes n'utilisent des sites factices dans l'unique but de contenir des citations pour mettre en valeur leurs publications. Ce système, qui peut paraître inégalitaire, est justifié par Sergei Brin de la façon suivante : « s*i je suis à la recherche d'un bon docteur dans la région, je vais regarder autour de moi et demander à mes amis de me recommander les bons docteurs. Ils pourront me désigner des personnes qui s'y connaissent mieux qu'eux – "ce gars connaît tous les docteurs de la Baie". J'irais ensuite voir cette personne pour la questionner*. » Sergei Brin s'appuie, pour justifier le fonctionenment du PageRank, sur le fait que dans la vie courante, quand nous avons besoin d'une information, nous la demandons à nos proches, qui nous dirigent eux-mêmes vers des personnes qu'ils jugent plus qualifiées pour nous conseiller, donc auxquelles ils ont confié une certaine autorité. Le PageRank donne donc la parole à tout le monde, mais se base ensuite sur un système de recommandations, car ses fondateurs estiment que certaines personnes sont plus dignes de confiance que d'autres dans un domaine spécifique. L'algorithme intègre donc également une dimension méritocratique, puisqu'il confère une autorité supérieure aux publications ayant été jugées sérieuses par d'autres.

La parole est donnée aux internautes, qui sont eux-mêmes responsables des résultats qu'ils obtiennent lorsqu'ils effectuent une recherche, puisqu'ils ont choisi de leur conférer de l'autorité en les recommandant. Google peut ainsi se présenter comme un médiateur extérieur totalement neutre, qui n'influe pas sur les résultats de recherche et s'assure seulement du bon fonctionnement de son algorithme. Le PageRank permet à Google de se réclamer d'objectivité puisqu'il n'intervient pas dans les résultats qu'il propose, et laisse l'ensemble des individus constituer les résultats. Or selon le principe même de la démocratie, qui fait relativement consensus dans nos sociétés, la décision du plus grand nombre constitue la meilleure décision possible : ce mode de fonctionnement permet donc à Google, en plus de limiter sa responsabilité quant aux résultats qu'il propose, de revendiquer fournir les meilleurs résultats possibles.

On notera toutefois la tension qui subsiste entre méritocratie et démocratie. Que serait une démocratie dans laquelle les votes n'ont pas tous le même poids ? Les fondateurs de Google dépassent cette apparente contradiction en faisant la distinction entre les internautes et leurs publications. Tous les internautes sont égaux puisqu'ils ont tous le droit de publier de façon égale, et donc de voter. Cependant toutes les pages publiées n'ont pas la même valeur, ce qui justifie le fait qu'elles n'aient pas toutes le même poids. De même dans notre système démocratique français, bien que chaque individu possède le droit de vote, tout le monde n'a pas la même autorité pour établir les lois, cette autorité ayant été confiée aux députés, qui se la sont eux-mêmes vue confiée par le vote du plus grand nombre. Certains individus ont donc dans notre société un poids plus important que les autres, du fait que le plus grand nombre le leur a conféré. Google cherche grâce à ce système d'autorité à juger un deuxième aspect

d'une page web : le premier étant sa pertinence dans le cadre de la recherche effectuée, et le second son importance (l'autorité qui lui a été conférée via les citations).

Comme le souligne James Grimmelmann : « *le génie de Google est que ses créateurs n'ont pas cherché à imposer un grand schéma organisationnel pour le web. À la place, ils ont demandé à tous les autres de le faire pour eux.* » Non seulement ce mode de fonctionnement a permis à Google de s'illustrer très rapidement comme étant le moteur de recherche fournissant, de très loin, les meilleurs résultats de recherche, puisqu'ils sont le reflet de l'autorité que les internautes ont conférée aux publications, mais il lui a également permis de se dédouaner de toute responsabilité, puisqu'il n'interfère pas dans le choix des résultats fournis. On notera d'ailleurs que Google refuse systématiquement et catégoriquement de modifier manuellement ses résultats de recherche, quand bien même il s'agisse de sites web xénophobes ou antisémites. James Grimmelmann nous donne l'exemple de la recherche du terme « *jew* » sur le moteur de reherche Google américain, qui fournissait en premier résultat, en 2004, un site antisémite. Cela avait été rendu possible grâce à une technique permettant de fausser les résultats de recherche et de se positionner artificiellement en première page, appelée GoogleBomb[17]. Suite à quoi un activiste juif a demandé à Google de supprimer ce site des résultats de recherche et de le remplacer par la page Wikipédia correspondant au mot « *jew* ». Google, malgré son refus d'interférer sur les résultats donnés par son algorithme, a toutefois, pour la première fois, pris l'initiative d'afficher un message en haut des résultats pour énoncer son désaccord quant aux idées véhiculées par ce site internet. Pour justifier cette non intervention, l'entreprise avance l'argument que son personnel ne fait pas entrer en compte ses opinions personnelles dans son travail, et qu'elles n'impactent donc pas les résultats de recherche. Cette réponse leur permet d'écarter leur responsabilité, et de dire de manière sous-entendue, selon les mots de James Grimmelmann : « *don't blame us, the computers did it* »[18], tout en conservant leur réputation d'objectivité et de non interférence sur les résultats de leur algorithme. Cette demi-intervention leur permettait ainsi de maintenir un *statu quo*, à savoir conserver la réputation de l'entreprise qui repose sur cette neutralité, ne pas prendre la responsabilité des résultats de recherche qu'elle fournit, et ne pas entacher son image. Cela est très révélateur du fait que les acteurs du Big Data sont mal à l'aise avec l'énorme pouvoir de captation qu'ils ont sur l'opinion publique. Ce refus n'est donc pas anodin, car il leur permet de conserver cette réputation de neutralité, et de limiter l'énorme responsabilité qu'ils ont à porter. Notons toutefois que Google s'efforce d'améliorer et de complexifier continuellement son algorithme pour éviter ce genre de détournements.

Enfin le dernier facteur revendiqué pour légitimer les algorithmes, et peut-être le plus fort d'entre eux, qu'il nous semble important d'aborder ici est pointé par Finn Brunton et Helen Nissenbaum, professeurs au département médias, culture et communication à l'Université de New York. Ils soulignent le fait que l'argument qui est continuellement donné aux individus sceptiques quant aux bienfaits apportés par le Big Data est qu'il nous permet d' « *obtenir ce*

[17] voir l'ouvrage *Google Dilemma* de James Grimmelmann (références dans la bibliographie) pour plus de détails sur le Google Bombing.
[18] « ne nous tenez pas pour responsables des résultats donnés par les ordinateurs »

qu'on veut avant même d'avoir su que nous le voulions »[19] [Brunton – Nissenbaum, 2011], et que cette idée constitue une forme de mythe sur les bienfaits apportés par le Big Data. Ce mythe reposerait selon eux sur la promesse d'une efficacité toujours plus importante des algorithmes du Big Data, et sur l'idée que nous sommes en train de vivre de profonds bouleversements sociaux au terme desquels la vie privée n'aura que peu d'importance, véhiculant ainsi l'idée que ceux qui y sont réfractaires ne sont plus dans l'ère du temps, ou bien qu'ils ont des choses à cacher. Selon les deux chercheurs, reprenant l'idée de Brian Pfaffenberger, professeur au département science, technologie et société à l'Université de Virginie[20], ce mythe que s'efforcent de construire les défenseurs du Big Data est calculé, car il est plus difficile d'argumenter contre les mythes et de faire appel à l'esprit critique de ceux qui y croient : « *myths are far harder to argue with and better at motivating people than a necessarily vague list of estimated future benefits*. »[21] [Brunton – Nissenbaum, 2011, p5] Finalement l'argument le plus fort, et celui revendiqué par la plupart des défenseurs du Big Data, est un argument qui ne fait pas appel à la rationalité des individus, qui se passe de toute réflexion et par là de toute remise en cause. C'est d'ailleurs là que réside sa force.

La légitimité du Big Data est toutefois souvent remise en question, et notamment l'algorithme de Google, loué par certains du fait que ce sont les internautes eux-mêmes, via les hyperliens, qui définissent la structure et la hiérarchie entre les publications sur internet, alors que d'autres, souligne Dominique Cardon, lui reprochent justement ce mode de fonctionnement reposant entièrement sur les individus. En effet selon eux, Google profite et exploite un travail qui n'est pas le sien. Google tirerait donc un profit économique du travail bénévole et désintéressé de l'ensemble des internautes, sans jamais leur en rétribuer les bénéfices. D'autres attaquent également l'algorithme sur le fait que toutes les pages n'ont pas le même poids grâce au système du PageRank, et l'accusent donc, comme nous l'avons énoncé précédemment, de ne pas être démocratique, ou du moins de n'être qu'une démocratie avortée. Toutefois, selon nous ces critiques ne pointent pas les vraies faiblesses du Big Data : dans le premier cas, on pourrait leur rétorquer que le bénéfice tiré par les utilisateurs est justement le service de qualité qui leur est offert, et pour la seconde critique, l'argument de Sergey Brin cité plus haut pour justifier le système méritocratique du PageRank nous semble cohérent. Du moins le postulat sur lequel il repose, selon lequel les pages publiées n'ont pas toutes la même valeur, nous semble difficilement discutable. Il est indéniable qu'un article de presse complet et citant ses sources a plus de valeur qu'un article de blog complotiste ne citant aucune source.

Cette façon de concevoir les algorithmes a permis aux acteurs du Big Data de prétendre que leurs algorithmes avaient résolu le « *problème fondamental du savoir humain* » [Gillespie, 2012] : ils permettent d'identifier les informations qui sont pertinentes pour les individus, en éliminant les biais inhérents à la manipulation et à la pensée humaines. Nous verrons

[19] « *You get what you want before you knew you wanted it* » - Finn Brunton & Helen Nissenbaum
[20] University of Virginia
[21] « *les mythes sont bien plus difficiles à remettre en question et bien plus efficaces pour motiver le public qu'une vague liste de potentiels futurs bénéfices* »

cependant que d'une part, les entreprises du Big Data ne respectent pas toujours les règles qu'elles se sont fixées et qu'elles revendiquent, et que d'autre part cette revendication d'objectivité et de neutralité est discutable.

Remise en question et limites

Nous avons énoncé précédemment que les acteurs du Big Data, et notamment Google, refusaient systématiquement de modifier manuellement leurs résultats de recherche, par souci de neutralité et de non implication dans la vie politique et éthique, comme dans le cas de la recherche sur le mot « *jew* », et afin de se détacher de toute responsabilité quant aux résultats que propose son moteur de recherche ; cela n'est toutefois pas complètement vrai, et il existe des antécédents pour lesquels les fournisseurs de moteurs de recherche ont été amenés à modifier leurs résultats. Par exemple en Europe, où l'incitation à la haine raciale est un point particulièrement sensible, selon l'exemple de James Grimmelmann, la justice française a sommé Yahoo![22] de supprimer de ses résultats tous les sites liés à la nostalgie nazie. Yahoo! a dans un premier temps refusé, mais a rapidement compris que cela pourrait entraîner le blocage de son moteur de recherche en France, ce qui aurait entraîné de lourdes pertes pour la société, et a donc obtempéré. Nous assistons donc ici à une censure, bien que potentiellement légitime, des résultats de recherche en France. De nombreux autres exemples existent de modifications des résultats de recherche suite à la demande des gouvernements, et nous nous attarderons particulièrement sur l'un d'entre eux, également énoncé par James Grimmelmann : en cherchant le mot « Tiananmen » sur le site chinois de Google, on s'aperçoit que les résultats ne montrent pas la célèbre photo de l'inconnu qui s'est dressé devant la ligne de tanks, symbole de la lutte non violente contre les systèmes autoritaires pour nous autres occidentaux, alors qu'elle apparaît parmi les premiers résultats sur les autres versions du moteur de recherche. Si la censure peut éventuellement se justifier dans le cas des sites internet nazis, elle devient beaucoup moins légitime dans le cas de la censure de manifestations pour la lutte pour la démocratie. Les moteurs de recherche sont le portail d'accès à l'information pour la très grande majorité des individus, et ils ont à cet égard une très lourd responsabilité à porter, bien qu'ils souhaitent s'en abstraire : ce sont eux qui rendent accessible ce que chacun juge comme étant la vérité à la plupart des individus sur la planète : les opinions des individus se forgent à partir des informations qu'ils trouvent sur internet au travers des moteurs de recherche. Quand ils acceptent de cacher aux internautes chinois les protestations à Tiananmen pour des raisons économiques, ils leurs offrent volontairement une vérité biaisée. Car il s'agit bien d'intérêts économiques, le retrait de certains résultats, dont celui de cette photo de Tiananmen, étant la condition requise par la dictature chinoise pour que les moteurs de recherches ne soient pas bloqués dans le pays. Cette construction de la vérité propre à chacun à travers les résultats fournis par les moteurs de recherche se constate d'ailleurs dans la réalité, car comme nous le fait remarquer James Grimmelmann, pour beaucoup de chinois Tiananmen est associé à 1949 et à la déclaration de la république populaire par Mao Zedong ; tenter désormais de rattacher Tiananmen à une manifestation populaire non-violente passerait pour beaucoup d'entre eux pour une tentative de manipulation de la part de l'occident. Mais au-delà de cette considération éthique et politique,

[22] Société américaine de services sur internet, comprenant notamment un moteur de recherche.

c'est ici toute l'idée d'objectivité sur laquelle repose pourtant essentiellement la confiance du public dans les acteurs du Big Data qui est remise en question. En effet les moteurs de recherche, en interférant manuellement sur les résultats de recherches fournis par leurs algorithmes perdent leur statut d'entité neutre, et par là objective.

James Grimmelmann relève également une autre source de biais : les acteurs du Big Data, et notamment Google, revendiquent le fait que leurs algorithmes fonctionnent de façon complètement automatisée, et se défendent ainsi de toute implication de leurs opinions personnelles ou d'autres sources de biais humain dans les résultats qu'ils fournissent. Mais quand bien même les résultats n'ont pas été modifiés manuellement, ces résultats ont été fournis par des algorithmes, et ces algorithmes ont été programmés par des ingénieurs. Un ordinateur est une instance neutre et n'effectue que les opérations qui lui ont été demandées. Les algorithmes ne font donc que ce que les ingénieurs leur ont commandé de faire, y compris les traitements automatiques qu'ils effectuent. Toute instance qui trouve à son origine une activité humaine ne peut pas se soustraire de toute source de biais, et encore moins de biais humain, bien que celui-ci soit indirect. Certes les algorithmes limitent les biais liés à l'action directe de l'homme, mais ils ne peuvent cependant pas être considérés comme totalement neutres et objectifs puisqu'à leur source on trouve tout de même une activité humaine. Il nous semble ici utile de rappeler ce que nous avons énoncé dans la première partie de cet exposé à propos de l'influence mutuelle entre les algorithmes et les individus : les acteurs du Big Data sont dépendants de l'opinion des utilisateurs, et cherchent donc à faire coïncider leurs algorithmes avec les aspirations des internautes. Par exemple si un résultat de recherche est estimé non pertinent par un grand nombre d'internautes, le fournisseur du moteur de recherche cherchera alors à modifier son algorithme de sorte que ce résultat n'apparaisse plus en première page. On constate ici le biais induit par le développement des algorithmes par des humains, qui l'orientent dans la direction qui leur convient, ou du moins qui converge avec les intérêts économiques de leur entreprise, à savoir la satisfaction des utilisateurs du service proposé. Les ingénieurs sont constamment en train de modifier les algorithmes, d'une part pour continuer à satisfaire les utilisateurs, mais aussi et surtout dans le but de contrer les individus qui contournent le fonctionnement des algorithmes pour qu'ils servent leur intérêt. C'est notamment le cas des experts en référencement pour les moteurs de recherche par exemple.

L'apparition des pratiques dites sociales sur le web ébranle de plus en plus le modèle de l'algorithme de Google, censé être fondé uniquement sur le dénombrement des citations, qui confèrent de l'autorité ou non aux publications. Ce modèle est en effet relativement élitiste, puisqu'il nécessite d'être un internaute « *publiant* » [Cardon, 13] pour pouvoir participer au classement des publications sur internet, alors que le modèle social permet à chaque internaute de devenir acteur en un simple clic (par exemple, un clic sur le bouton « *like* » sur le réseau social Facebook permet à l'internaute de manifester son intérêt pour une publication). L'introduction de la considération de ce qu'on appelle l' « *affinité* » [Cardon, 2013], promue par les réseaux sociaux à travers les « *like* » dans le cas de Facebook et les « *+1* » dans le cas de Google pour ne citer que les plus gros, ont donc fait apparaître de

nouvelles variables à prendre en compte pour la priorisation des résultats, qui vont à l'encontre du modèle algorithmique initial de Google.

Dans l'exemple cité précédemment, on assiste à une modification de la façon dont un algorithme est codé afin qu'il ne donne plus les mêmes résultats, mais d'autres formes d'interventions encore plus directes sur les algorithmes existent. En effet, comme nous le fait remarquer James Grimmelmann, il arrive par exemple parfois que Google intervienne directement sur les PageRank de certains sites internet (comme dans les cas de censures cités précédemment sauf que dans ces cas-là les entreprises faisaient écho à une décision de justice), afin de les dévaluer, les faisant ainsi disparaître des résultats de recherche. Cela s'explique par le fait que certains internautes malintentionnés ont compris qu'en créant un certain nombre de sites internet se renvoyant les uns vers les autres permettait d'augmenter artificiellement leurs PageRank respectifs, et ainsi d'améliorer leurs positions dans les résultats de recherche. Cette méthode, appelée un *link farm*, a rapidement été remarquée par Google, qui a interdit son usage et a décidé de diminuer manuellement le PageRank de ces fermes de liens. Ici l'intervention semble légitime, puisque par exemple si un individu fait une recherche pour trouver un manteau, il cherche à accéder à un vendeur de manteaux recommandé par les autres internautes, et non pas à un vendeur dont la notoriété a été artificiellement augmentée grâce à une ferme de liens, mais le biais introduit n'en subsiste pas moins pour autant. Nous avons donc ici affaire à un nouvel exemple d'implication des entreprises du Big Data sur les résultats fournis par leurs algorithmes. Ces interventions sont d'autant plus néfastes pour la confiance des internautes dans les algorithmes du fait qu'elles ne leur sont pas imposées par une décision de justice. Il devient alors difficile de situer la limite, et l'individu peut être amené à penser que les entreprises sont susceptibles d'interférer de façon arbitraire sur leurs résultats, ou du moins en suivant des motivations qui ne vont pas dans leur intérêt.

Le principal problème pour les acteurs du Big Data, comme le souligne à juste titre Dominique Cardon, vient donc du fait que certains internautes comprennent les modes de fonctionnement des algorithmes et agissent en fonction d'eux, afin d'exploiter au mieux leurs failles pour servir leurs propres intérêts ; chaque internaute le fait d'ailleurs inconsciemment de façon plus ou moins poussée : lorsqu'un individu utilise un hashtag[23] fortement utilisé sur Twitter afin de donner de la visibilité à son *tweet*, alors que le hashtag n'est absolument pas représentatif du contenu de son message, il exploite le mode de fonctionnement de l'algorithme pour servir son intérêt, à savoir dans ce cas la volonté de donner de la visibilité à sa publication. Cela oblige les fournisseurs de ces services à identifier ces failles et à agir pour rectifier les problèmes qu'elles posent, notamment dans le cas des moteurs de recherche que nous avons cité précédemment, cassant ainsi la position de neutralité et d'extériorité qu'ils s'efforcent de construire et de maintenir. De même, comme le remarque Dominique Cardon, le cas de Google est très révélateur : lui qui s'efforce et se défend de seulement dénombrer des

[23] Un hashtag permet de marquer un contenu avec un mot-clé plus ou moins partagé. Il s'agit d'un mot précédé d'un « # ».

liens sans s'intéresser au contenu, est de plus en plus poussé et forcé à porter un jugement sur la nature des hyperliens et donc sur la qualité du contenu vers lequel ils mènent.

Une autre critique qui est faite aux moteurs de recherche, et notamment à Google, est que leurs algorithmes n'offrent finalement de visibilité principalement qu'à un très petit nombre d'institutions à forte renommée, qui monopolisent la majeure partie de l'espace sur le web. Dans le cas de Google et de son PageRank, ce phénomène peut s'expliquer facilement : toutes les publications ont en effet potentiellement la même possibilité de recevoir des liens de la part des autres, mais dans la réalité, nous dit Dominique Cardon, « *un très petit nombre de pages attire un nombre considérable de liens, pendant que la très grande majorité des sites sont liés à très peu de sites et ne sont souvent cités par aucun* » [Cardon, 2013, p88]. Les principaux acteurs du web, et les institutions qui disposent déjà d'une forte popularité à l'extérieur du web du fait de leur notoriété, obtiennent rapidement un PageRank élevé, et donc une position prédominante. Cette position prédominante attire à son tour de nombreuses citations, renforçant ainsi leur position. Ce phénomène est par ailleurs renforcé par le fait que les sites internet ont tendance à ne citer que des « *sites ayant une autorité égale ou supérieure à la leur* ». [Cardon, 2013, p89] A titre d'exemple, prenons l'exemple du site internet de la marque à forte renommée Porsche : il existe de nombreux sites internet qui parlent de voitures de cette marque, et quand ceux-ci veulent faire référence à un site internet pour appuyer leurs propos, ils auront souvent tendance à citer celui du constructeur, qui bénéficie d'une forte renommée et constitue un gage de fiabilité. Il résulte de cet état de choses que les principaux sites internet (ceux des grands acteurs du web ou des grandes institutions internationales), perçoivent autant de citations vers leurs publications que de clics de la part des internautes, sans qu'on ait la possibilité de savoir qui de *l'autorité* ou de la *popularité* a préexisté à l'autre. On voit donc ici que l'algorithme de Google, revendiqué par ses instigateurs comme mettant en place un système méritocratique absent de tout biais humain, se retrouve finalement à renforcer la notoriété de ceux qui en disposaient déjà le plus. Au-delà du problème que cela pose par rapport à ce dont se revendique Google (un système méritocratique sans failles), qui concerne plutôt les auteurs de publications, on est ici amené à constater que les individus sont finalement enfermés dans une bulle dans laquelle on ne leur rend visible qu'une infime minorité de l'information disponible sur internet. Les acteurs du Big Data, qui se revendiquent de neutralité et d'objectivité, ne nous offrent finalement qu'une *vérité* étroite et peu diversifiée, qui n'est jamais remise en question.

Enfin, il nous semble important de questionner, comme nous le suggèrent Finn Brunton et Helen Nissenbaum, la nécessité d'une telle collecte des données personnelles pour alimenter les algorithmes. Même en admettant que les algorithmes du Big Data rendent un certain nombre de services appréciables aux individus, et que pour fonctionner et être efficaces ils doivent être alimentés en données, le degré à laquelle est pratiquée la collecte des données personnelles est beaucoup trop important pour être justifié. Par exemple, l'utilisation du profilage afin de proposer des publicités ciblées peut-être acceptable, mais ce profilage est d'une telle ampleur que cette seule fin ne peut le justifier à elle seule. Toutes nos données de navigation, toutes les traces que nous laissons sur notre passage (et nous en laissons au

minimum à chaque clic que nous effectuons), sont collectées et stockées par défaut, et ce même si les entreprises qui les collectent n'y voient pas d'intérêt immédiat : les données sont donc collectées la plupart du temps avec une absence d'intention au départ, sans objectif préalablement défini. Stocker les données coûte peu cher, et les entreprises du Big Data savent que potentiellement, tôt ou tard leurs algorithmes pourront exploiter ces données initialement inutiles, et en extraire des informations à valeur économique positive. L'absence d'intention n'est d'ailleurs pas seulement le privilège de ceux qui collectent les traces, il en est de même pour l'émetteur de la trace, comme le rapporte Louise Merzeau, enseignante-chercheuse en sciences de l'information et de la communication, en citant Sybille Krämer : l'internaute laisse des traces sans en avoir l'intention, il n'en est pas l'instigateur et n'en a souvent même pas conscience, c'est justement « *ce qui échappe à notre attention, à notre contrôle ou à notre vigilance qui, à partir de nos actes, prend la forme d'une trace* » [Merzeau, 2013, p8]

Cette réflexion sur la légitimité des algorithmes nous amène donc d'une part à remettre en question la notion d'objectivité des algorithmes, revendiquée par tous les acteurs du Big Data, et sur laquelle repose pour une très grande partie leur acceptation par le public, mais aussi à douter du bien-fondé de ces algorithmes et de l'alimentation en données qu'ils requièrent, qui ne nous permettraient d'avoir accès qu'à une vérité tronquée, et extrêmement fermée. L'idée d'internet comme étant un univers démocratique et égalitaire, très souvent reprise par les défenseurs du Big Data, serait également un leurre, puisque ce qu'on en a fait, via le Big Data, serait en fait tout le contraire. Nous consacrerons la prochaine partie de cet exposé à identifier les voies possibles pour se prévenir des actions néfastes du Big Data sur notre utilisation du web ; nous tâcherons par ailleurs de déterminer si ces champs d'actions sont eux mêmes légitimes et ne posent pas de questions éthiques.

Protection de la vie privée

Nous avons montré dans la première partie que le Big Data capte de plus en plus de données provenant de toutes sortes de capteurs d'origines toujours plus diverses (téléphones, objets connectés, réseaux de communication, …), qui concernent des aspects toujours plus privés de notre vie. Par ailleurs la combinaison de ces données entre elles et leur analyse grâce aux algorithmes de prédiction peuvent révéler des aspects intimes de notre personnalité, même si la nature des données utilisées pour alimenter les algorithmes parait peu personnelle. Les acteurs du Big Data légitiment cette captation par défaut des données par le fait que leurs algorithmes se veulent objectifs, et qu'ils permettent ainsi aux individus d'évoluer dans un internet démocratique et égalitaire, dans lequel chacun peut s'exprimer et faire entendre sa voix. Or comme nous l'avons montré dans la seconde partie, cette objectivité se révèle être un leurre, et le régime qu'elle instaure ne serait qu'un subside de démocratie, favorisant finalement les puissants. Toutefois, il existe un relatif consensus quant au fait que, dans certains cas au moins, le Big Data peut se révéler utile et bénéfique. Nous faisons donc face ici à un paradoxe : les individus exigent de profiter des bénéfices apportés par le Big Data, tout en souhaitant se protéger de la collecte des données et de l'intrusion dans la vie privée sur lesquelles reposent justement le Big Data : comment alors concilier les deux ?

Dans cette dernière partie, nous tâcherons d'identifier dans un premier temps les moyens qui s'offrent aux internautes pour lutter et se protéger contre cette captation des données, de manière individuelle ou collective. Nous chercherons ensuite à savoir si ces moyens sont éthiquement acceptables et quelles sont leurs limites.

Méthodes d'obfuscation

Une communauté de plus en plus importante sur le web s'est rassemblée pour mettre en place des stratégies de lutte contre la collecte et l'exploitation des données personnelles. Ces pratiques, selon Finn Brunton et Helen Nissenbaum, sont dites d'« *obfuscation* » [Brunton – Nissenbaum, 2011]. Il existe différentes formes d'obfuscation, qu'il convient de distinguer, et dont nous tenterons de donner un aperçu, en nous basant sur celles relevées par Finn Brunton et Helen Hessenbaum. Il est nécessaire, avant tout, de distinguer les méthodes individuelles, qui peuvent être effectuées par un internaute isolé, des méthodes collectives, qui s'inscrivent nécessairement dans un groupe d'internautes, et voient d'ailleurs, comme nous le verrons par la suite, leur efficacité augmentée à mesure que le nombre d'individus s'inscrivant dans le groupe augmente..

Parmi les méthodes individuelles, les premières que nous allons aborder, dites « *time-based* », ne sont valables, comme leur nom l'indique, que dans les cas où le temps est limité, et non pas pour insérer un doute de façon permanente. Ces méthodes consistent à multiplier le nombre de données plausibles disponibles, en dissimulant la vraie parmi elles. Cette méthode n'est pas permanente, car le temps permettrait aux algorithmes de l'extraire, mais elle est valable quand le temps imparti est court. Les deux enseignants nous donnent l'exemple d'une pratique qui a eu lieu en 2010 dans le domaine du trading haute fréquence, qui consiste à envoyer un très grand nombre d'ordres tout à fait inutiles, dans le but de forcer les concurrents à les analyser, ce qui les ralentit et permet au système émetteur, qui lui les ignore, de gagner un temps significatif. En effet le trading haute fréquence consiste à exécuter à très grande vitesses des transactions financières via des algorithmes informatiques, donc chaque milliseconde compte, et la lecture de quelques milliers de transactions factices peut faire perdre un temps non-négligeable à cette échelle.

Il existe des méthodes d'obfuscation plus pérennes dans le temps, et qui ne peuvent pas être opérées par un individu isolé, comme c'est le cas pour l'exemple précédent, mais nécessitent une action collective ; pour ce genre d'obfuscations, dites obfuscations collectives, qui fonctionne selon un mécanisme itératif : plus le nombre d'individus impliqués est important et plus leur action sera efficace. Ici nous citerons un exemple de situation réelle, relevé par Finn Brunton et Helen Hessenbaum, afin d'illustrer l'idée : au cours de la seconde guerre mondiale, une grande partie de la population danoise s'est mise à porter l'étoile jaune afin d'empêcher les allemands de trouver les juifs ; cet exemple illustre parfaitement l'idée d'obfuscation collective, puisque le port de l'étoile n'aurait eu aucun sens si un seul individu s'en était paré, et parallèlement plus le nombre d'individus la portant est important, et plus la tentative de dissimulation est efficace. Pour nous recentrer sur le web, nous prendrons l'exemple de Tor, qui est un système permettant d'utiliser et d'échanger sur internet de façon complètement anonyme. Le fonctionnement est simple, sur ce réseau les messages ne sont pas passés directement de l'émetteur au destinataire, mais passent par un nombre aléatoire de

relais, sans que l'on sache jamais qui en est l'émetteur ni le destinataire. Lorsqu'on reçoit un paquet, on a uniquement l'information sur le relai qui nous l'a passé, qui peut aussi bien être l'émetteur ou quelqu'un d'autre, et l'information nous permettant de savoir si on est le destinataire du message ou si l'on doit le transmettre aléatoirement à un autre relai. De la même manière que dans l'exemple danois, le mécanisme est ici aussi itératif : plus le nombre de participants au réseau sera important, plus le nombre de relais le sera, et donc plus l'obfuscation sera efficace et moins les messages seront traçables.

Ces deux premiers types d'obfuscation permettent de masquer complètement ses traces, mais elles ne permettent par conséquent pas d'interagir avec des sociétés qui offrent des services comme les réseaux sociaux, les plateformes de vente en ligne, … Une solution alternative commence donc à voir le jour, appelée obfuscation sélective. Ces méthodes permettent de sélectionner les informations qu'on veut fournir à ces services, et de choisir celles qu'on veut leur cacher, qui seront donc cryptées puis stockées sur des serveurs alternatifs. Elles seront alors uniquement visibles par les utilisateurs avec lesquels on choisit de les partager. Finn Brunton et Helen Hessenbaum donnent ici l'exemple d'une extension du navigateur Mozilla Firefox, appelée FaceCloak, et qui permet de choisir quelles informations on souhaite partager avec Facebook, et qui seront donc envoyées sur leurs serveurs, et lesquelles on souhaite leur cacher. Celles qui leur sont cachées seront donc uniquement accessibles aux utilisateurs avec lesquels on choisit de les partager, car ils seront les seuls à pouvoir les décrypter. Néanmoins ces solutions sont contraignantes dans le sens où leur utilisation n'est pas transparente et nécessite un effort plus ou moins important de la part des internautes. De même l'étendue de leurs possibilités sur le web se trouvent parfois amoindries. Par exemple, dans le cas de FaceCloak, son utilisation requiert que les utilisateurs avec lesquels on souhaite partager des données qui ne le sont pas avec Facebook, possèdent également le *plugin*. Mais il s'agit tout de même d'une alternative permettant de conserver la propriété de ses données en permettant leur utilisation uniquement pour l'usage auquel on les destine, tout en étant en capacité de les partager, ce qui s'avère presque indispensable à l'heure du web dit « social ».

Enfin le dernier type d'obfuscation identifié par Finn Brunton et Helen Nissenbaum consiste à rendre les données transmises ambigües et peu fiables, et ce de façon permanente, contrairement à l'obfuscation basée sur la contrainte de temps. Cette méthode est notamment utilisée pour anonymiser les requêtes sur les moteurs de recherche, mais également sur le réseau BitTorrent. Ici les deux enseignants nous donnent l'exemple d'une extension pour navigateurs internet appelée TrackMeNot, et qui permet d'envoyer un certain nombre de requêtes de façon aléatoire à chaque fois qu'on émet une recherche. Ainsi la recherche qu'on effectue est dissimulée parmi les autres. De plus on choisit les sources dans lesquelles puise l'extension afin de générer ces recherches. L'extension tâche également d'imiter les recherches de l'utilisateur afin d'améliorer son efficacité et de rendre les requêtes générées automatiquement moins détectables.

Nous avons donc décrit dans cette partie la grande variété de méthodes d'obfuscation, qui sont plus ou moins pertinentes en fonction de l'usage qu'on veut en faire. On peut facilement

imaginer que ce nombre va être amené à s'accroître, car les usages et les pratiques sur internet évoluent et augmentent constamment. L'apparition de ces pratiques d'obfuscation s'inscrit en réponse à la montée en puissance des entreprises du Big Data, et pour rétablir le rapport de force très inégal qu'elles entretiennent avec les individus. Nous verrons donc dans la partie suivante dans quelles mesures elles peuvent se justifier.

Justifications et limites

La montée en puissance du Big Data, et ce notamment dans le secteur privé, pose de nouvelles problématiques dans le cadre de la protection de la vie privée. Kate Crawford et Jason Schultz nous font d'ailleurs remarquer que les moyens actuels mis en œuvre pour la protection de la vie privée sont très largement dépassés et insuffisants pour répondre aux risques auxquels sont exposés les individus à l'ère du Big Data. La combinaison et l'analyse de sources hétérogènes de connaissance grâce aux algorithmes prédictifs augmentent drastiquement les données qui doivent désormais être considérées comme privées, car même si elles paraissent peu intrusives en apparence, elles peuvent révéler des aspects très intimes de la vie des individus une fois combinées. Les algorithmes prédictifs peuvent être utilisés pour anticiper les inclinations personnelles des individus, dans le but de leur soumettre une offre commerciale adaptée ; les entreprises peuvent ainsi, par exemple, adresser aux individus des publicités personnalisées, et ce à partir de la connaissance approfondie de leur vie privée, déduite par les algorithmes prédictifs. Mais au-delà des questions éthiques que pose le Big Data et que nous avons déjà abordées dans les parties précédentes, de nombreux enjeux juridiques se font jour. Le Conseil d'Etat français s'est d'ailleurs récemment emparé du problème dans un rapport portant sur « Le numérique et les droits fondamentaux »[24]. Ce rapport s'est donné pour objectif de déterminer l'influence des algorithmes sur les individus et les risques qui y sont liés, et d'encadrer leur utilisation. Dans ses recommandations pour encadrer le Big Data, le Conseil d'Etat identifie plusieurs pistes, et notamment la transparence des données personnelles utilisées par les algorithmes afin que les individus puissent avoir un regard dessus et faire valoir d'éventuelles critiques, et le renforcement des actions pour détecter les discriminations opérées par les algorithmes prédictifs. Mais la détection des discriminations est très complexe car les algorithmes ne sont pas prévisibles, et leurs effets ne sont d'ailleurs souvent même pas complètement compris par les ingénieurs qui les programment. Kate Crawford et Jason Schultz ont d'ailleurs très clairement identifié ce problème, et nous font remarquer que la régulation de la collecte d'informations sensibles est déjà en elle-même extrêmement complexe, mais que réguler une forme d'analyse l'est encore plus : comment prévoir le fait qu'un algorithme va fournir une information qui va à l'encontre du respect de la vie privée à partir d'une donnée en apparence inoffensive ? Les deux chercheurs encouragent donc les gouvernements à contrôler l'intention des processus d'analyse du Big Data, et la fin visée par le processus d'analyse, plutôt que la collecte des données elle-même : dans le cas où l'utilisation du Big Data mène à une discrimination, les individus qui en sont victimes pourraient alors saisir la justice. Kate Crawford et Jason Schultz donnent pour exemple le cas d'un employeur et d'un candidat pour un poste à pourvoir : « *if a potential employer used Big Data to predict how honest certain job*

[24] Rapport datant de 2014 et accessible à cette adresse : http://www.alain-bensoussan.com/wp-content/uploads/2014/10/28040001.pdf

applicants might be, these applicants could then exercise their data due process rights »[25] [Crawford - Schultz, 2014, p109]. Selon l'idée des deux chercheurs, dans le cas où l'utilisation du Big Data par un employeur a pour conséquence la discrimination de certains candidats, donc une fin répréhensible par la loi, le candidat serait en mesure de saisir la justice pour protéger ses droits. On ne questionne pas ici le fonctionnement de l'algorithme utilisé, ni les données qui lui ont été fournies, mais plutôt le but dans lequel il a été sollicité ; on remet ici en quelque sortes l'humain au centre de la préoccupation de la justice, puisque ce sont les intentions de celui qui utilise un algorithme qui sont jugées, et non pas cet algorithme en lui-même. Cette autre piste pour encadrer juridiquement le Big Data consiste à le remettre au service de l'individu. Elle est d'ailleurs abordée dans le rapport rédigé par le Conseil d'Etat qui reconnait que l'usage des données personnelles par le Big Data permet tout de même d'optimiser de nombreux services et rend ainsi service à l'individu, et préconise donc de l'exploiter pour améliorer et rendre plus performantes les politiques de santé, d'éducation, de promotion de la culture, ou encore pour simplifier les démarches administratives ; le Conseil d'Etat préconise de mettre le Big Data au service de l'intérêt général. Kate Crawford et Jason Schultz vont toutefois beaucoup plus loin dans leur réflexion, et identifient trois droits fondamentaux des individus qui peuvent être altérés par le Big Data et doivent être protégés : la vie, la liberté et la propriété. Ces trois aspects ont selon eux tous trait à la sphère privée, et peuvent être affectés par le Big Data. La liberté, en tant que « the *right to be let alone* »[26] [Crawford – Schultz, 2014, p111], c'est-à-dire le droit à la solitude et au libre arbitre, correspond au respect de la vie privée, dans le sens où empêcher l'individu d'être seul constitue une atteinte à son intimité. Elle est très fortement atteinte par les entreprises du Big Data, qui enferment l'internaute leur environnement, selon les règles qu'ils ont eux-mêmes fixées ; or comme nous le verrons ensuite, s'affranchir des services de ces entreprises a un coût de plus en plus important pour l'individu. De même, les algorithmes prédictifs, à partir de nos données personnelles, peuvent être utilisés pour décider d'accorder ou non un prêt, ce qui influe par conséquent sur le droit de propriété. La vie des individus est quant à elle continuellement sollicitée et impactée par le Big Data, via nos données privées, puisqu'il a des implications sur les questions de sécurité par exemple, ainsi que sur tous les aspects que nous avons énoncés au cours de cet exposé. Selon eux, plus le Big Data impacte sur un aspect important de notre vie privée, et plus on doit questionner le moyen par lequel il est arrivé à ce résultat. Par exemple, les questions de santé devront faire l'objet de beaucoup plus d'attention que les problèmes de publicité ciblée car les informations relatives à la santé des individus sont beaucoup plus sensibles et intimes. Quand un usage est concerné par deux catégories, on doit appliquer le niveau de contrôle le plus élevé. Pour illustrer cette dernière idée, on peut imaginer le cas de l'utilisation d'algorithmes prédictifs pour obtenir des informations relatives à la santé des individus, dans le but de cibler uniquement ceux qui sont jugés en bonne santé et leur proposer des offres pour une mutuelle. On a donc ici affaire à un cas concerné par les catégories santé et publicité : on doit appliquer ici, selon Jason Schultz et Kate Crawford le niveau de contrôlé réservé à la santé. Selon les deux chercheurs, il est bien évidemment nécessaire de s'assurer que les décisions prises par le Big Data apparaissent comme équitables

[25] « Si un employeur potentiel utilisait le Big Data pour prédire l'honnêteté des postulants à un poste qu'il pourvoit, alors ces postulants pourraient saisir la justice. »
[26] « *Le droit à être laissé tranquille* »

et justes, de même que les processus qui mènent à ces décisions doivent être complètement transparents et très contrôlés, afin justement d'être en mesure de vérifier l'équité de ces décisions. Mais il faut surtout s'assurer que la fin visée par l'entité qui sollicité les algorithmes respecte la vie privée et les droits des individus.

Nous avons donc vu dans un premier temps des pistes pour faire face aux enjeux juridiques impliqués par le Big Data et les profonds bouleversements qu'il a apportés concernant les problématiques de protection de la vie privée. Toutefois, les gouvernements peinent à saisir les enjeux liés au Big Data quant au respect de la vie privée des individus, et sont victimes des lobbies très actifs des entreprises du secteur. Or tant que les gouvernements n'agissent pas, les individus restent vulnérables et ne peuvent contrôler quand les données les concernant sont collectées, ni les usages qui en sont faits. C'est pourquoi les méthodes d'obfuscation, énoncées en première partie, apparaissent pour beaucoup comme une alternative à l'inaction des gouvernements. Finn Brunton et Helen Nissenbaum définissent l'obfuscation de la façon suivante : « *producing misleading, false, or ambiguous data to make data gathering less reliable and therefore less valuable* »[27]. Il faut donc distinguer l'obfuscation d'autres pratiques comme la dissimulation (grâce aux méthodes de cryptographie par exemple) ou l'effacement. Dans le cas de l'obfuscation, il s'agit de rendre les données moins lisibles, d'y ajouter du bruit, des perturbations, afin qu'elles soient plus difficilement exploitables par les algorithmes, ou de les induire en erreur. On peut associer cette idée à celle d'un canal de communication, comme le réseau téléphonique : parfois la communication est mauvaise et on entend de forts grésillements, qui rendent les échanges difficilement compréhensibles ; l'obfuscation consisterait à émettre volontairement ces parasites, qui apportent du bruit à la communication, afin de conserver sa conversation confidentielle. Il s'agit de faire en sorte que les messages soient reçus avec une partie du contenu manquante, ou même qu'ils aient été suffisamment altérés pour qu'ils deviennent difficiles à reconstituer, voire même que leur sens ait été complètement changé. Il existe deux types de moyens d'actions : les premiers, dits de « *désinformation destructrice* » [Brunton – Nissenbaum, 2011] consistent à détruire une partie de l'information avant qu'elle n'arrive entre les mains des acteurs du Big Data ; les seconds, plus subtiles, et appelés « *désinformation* constructrice », consistent non pas à détruire de l'information, mais à en remplacer des parties afin d'en changer le sens. Les méthodes destructrices ont rapidement montré leurs limites puisqu'elles sont facilement détectées par le récepteur, qui peut alors prendre des mesures pour contrer cette destruction d'information ; pour reprendre l'exemple de la communication téléphonique, si le destinataire ne reçoit qu'un mot sur trois, le message ne sera plus porteur de sens et il s'en apercevra naturellement. En revanche dans le cas des secondes, le récepteur ne sait pas que le sens du message a été transformé, puisqu'il reçoit toujours un message complet, sans anomalies et porteur de sens. L'obfuscation consiste avant tout à augmenter le coût nécessaire à l'exploitation des données pour ceux qui souhaitent les collecter, c'est pourquoi on considère que c'est une méthode plus faible que celles comme la cryptographie, dont la sécurité et la sureté est mesurable et dépend de paramètres définis. Cependant, la plupart du temps les méthodes fortes ne sont pas autorisées par la plupart des sociétés qui offrent une expérience personnalisée et dont le

[27] « *Produire des données trompeuses, fausses ou ambigues pour rendre la collecte des données opérée par le Big Data moins fiable, et ainsi moins précieuse.* »

modèle économique repose sur les pratiques du Big Data ; l'obfuscation se présente donc, dans ces cas particuliers, comme la seule alternative possible.

Ces pratiques sont justifiées par leurs auteurs par le fait que les individus laissent des traces et sont profilés au cours de toutes leurs moindres transactions sur le web (le moindre clic est collecté puis analysé), et ce souvent sans même qu'on les en informe, et encore moins qu'on leur laisse le choix. L'obfuscation constitue donc une forme de résistance à cette surveillance continue qui est imposée aux individus. Les sociétés de service du Big Data, elles, affirment que les individus ont le choix de ne pas utiliser leurs services s'ils sont opposés à la collecte de leurs données personnelles. Mais Finn Brunton et Helen Nissenbaum, ainsi que les défenseurs de l'obfuscation, dénoncent le fait que bien que l'usage de ces services ne soit pas obligatoire, choisir de s'en priver présente d'ores et déjà un coût social certain, et celui-ci ne fait que s'amplifier avec le temps ; les individus seraient en quelque sorte prisonniers, et ce de plus en plus fortement, des services offerts par les sociétés du Big Data, sous peine de se retrouver en marge et isolé. Ce problème rejoint ce que Durkheim appelle le *déterminisme social* : on ne choisit pas notre langue maternelle, et bien qu'on soit libre de l'abandonner, le coût social que représenterait cet abandon est tel qu'il suffit à nous en dissuader. De même, Finn Brunton et Helen Nissenbaum estiment que les enjeux sont si importants et le stockage des traces laissées sur internet déjà tant systématisé, que nous ne pouvons pas nous permettre d'attendre que la loi évolue et encadre les pratiques de collecte et d'utilisation des données personnelles. D'autant plus que même concernant les données inexploitables à l'heure actuelle, apparemment inoffensives et non intrusives, un sens et un usage pourront très probablement leur être assignés à l'avenir, avec l'évolution et l'amélioration des méthodes d'analyse et de corrélation des algorithmes ; il serait donc de notre devoir d'agir dès maintenant, si ce n'est dans notre intérêt actuel, au moins dans le but de se protéger demain. L'obfuscation n'est donc pas selon eux un moyen pérenne pour se protéger contre les excès de la collecte systématique des données personnelles, mais plutôt une pratique temporaire, destinée à compenser les retards de nos sociétés actuelles en termes de législation dans ce domaine, avant que ceux-ci ne soient comblés. Il s'agit de préserver les droits essentiels des individus et rééquilibrer le rapport de forces entre les sociétés faisant usage du Big Data et les individus qui en sont victimes, qui va aujourd'hui très largement en la défaveur des seconds. En effet, les individus disposent d'un pouvoir d'action et d'une capacité de connaissance beaucoup moins forts que ceux des acteurs du Big Data : ils ne savent pas exactement quelles données sont collectées, ni à quel moment elles le sont, à qui elles sont partagées et dans quel but.

Bien que ce constat d'inégalité du rapport de forces entre internautes et sociétés du Big Data soit difficilement contestable, il n'en reste pas moins que l'obfuscation repose souvent sur le mensonge. Il s'agit en effet la plupart du temps de faire passer des informations erronées pour des informations non altérées, c'est pourquoi nous interrogerons dans la partie suivante les limites éthiques de ces pratiques.

Obfuscation et éthique

Comme le rapportent Finn Bruton et Helen Nissenbaum, ses pourfendeurs affirment que l'obfuscation est une pratique malhonnête qui repose sur le mensonge, car elle consiste à

fournir des informations fausses ou altérées ; la cryptographie serait selon ce point de vue une méthode préférable puisqu'on empêche tout simplement la lecture des informations plutôt que les falsifier. Certains défenseurs de l'obfuscation se défendent en affirmant qu'ils fournissent des informations ambigües, et non pas des données erronées comme on le leur reproche. Mais nous retiendrons surtout l'argument apporté par les deux chercheurs, qui relèvent que seul le philosophe allemand Emmanuel Kant affirme que le mensonge n'est jamais juste et toujours moralement répréhensible, ne serait-ce que parce qu'il nuit à l'autre, dans le sens où il l'empêche d'agir rationnellement et en toute connaissance de cause, puisque des choses lui sont cachées ; selon lui, on doit toujours dire la vérité, quel qu'en soit le prix, et ce y compris si cela mène à un meurtre injustifié. Un exemple extrême est d'ailleurs souvent utilisé pour illustrer le propos du philosophe : dans le cas où un criminel me somme de lui fournir une information qui met en danger ma propre vie ou la vie d'un autre individu, il estime que je dois tout de même fournir cette information, sans quoi je commettrais une injustice envers la morale, et commettre une telle injustice correspond selon lui à commettre une injustice envers tout le reste de l'humanité, puisque l'humanité repose sur la morale qu'elle a fondée. Excepté ce cas extrême, il existe un relatif consensus de la part des philosophes, selon lequel le mensonge serait tolérable dans certains cas : l'individu doit-alors s'interroger pour juger si le bénéfice du mensonge est plus important que son coût. Nous remarquerons que cette pensée est problématique, dans le sens où les notions de coût et de bénéfice sont souvent relatives aux individus : ce qui est bien pour moi n'est pas forcément bien pour les autres. Le cas de la seconde guerre mondiale et de la déportation des juifs apporte néanmoins de nos jours un exemple fort, et peu se risquent à affirmer qu'il était immoral de mentir pour protéger des individus menacés de déportation. De manière plus théorique, nous reprendrons la pensée de Benjamin Constant, homme politique français ayant vécu au XIXème siècle, qui s'oppose à celle de Kant :

« Dire la vérité est un devoir. Qu'est-ce qu'un devoir ? L'idée de devoir est inséparable de celle de droits : un devoir est ce qui, dans un être, correspond aux droits d'un autre. Là où il n'y a pas de droits, il n'y a pas de devoirs. Dire la vérité n'est donc un devoir qu'envers ceux qui ont droit à la vérité. Or nul homme n'a droit à la vérité qui nuit à autrui. »[28] [Constant, 1797]

Selon Constant, chaque homme dispose d'un droit inaliénable qui est celui de conserver sa dignité. Il ne peut donc pas exister de devoir qui puisse nuire à l'autre, car nuire à l'autre consiste à nuire à sa dignité. Pour reprendre l'exemple de la seconde guerre mondiale, les individus n'avaient selon lui aucun devoir de vérité envers les criminels nazis, puisque la leur dire et dénoncer des individus recherchés aurait nuit à la dignité de ces individus. De même concernant le Big Data, si un individu considère que les entreprises qui collectent ses données lui nuisent, il n'a pas de devoir de vérité envers elles. La question reste de savoir si ces entreprises lui nuisent plus en collectant ses données, que lui en les leur cachant.

Un autre reproche qui est régulièrement fait à l'obfuscation, là encore relevé par Finn Brunton et Helen Hessenbaum, réside dans le fait que le principe auquel elle est liée consiste à diriger

[28] Citation tirée de l'ouvrage : *Des réactions politiques*, de Benjamin Constant.

les sociétés de collecte et d'analyse des données vers des cibles qui ne protègent pas leurs données, et sont par conséquent jugées plus faciles. Les sociétés sont intéressées par la collecte de données en raison de son faible coût, il s'agit donc, pour l'obfuscateur, d'augmenter le coût de collecte de ses données afin de les dissuader. Le problème moral ici est que l'individu qui protège ses données profite des bénéfices et des services rendus par les sociétés du Big Data, sans pour autant contribuer au coût de ces structures, le faisant ainsi reposer sur les autres utilisateurs. L'individu se protège donc lui-même en faisant subir le préjudice auquel il se juge exposé au reste de la communauté. Si nous prenons l'exemple d'Adblock Plus, la célèbre extension disponible sur les navigateurs Google Chrome et Mozilla Firefox permettant de bloquer l'affichage des publicités lors de la navigation sur internet, l'individu qui l'utilise continue de profiter du contenu des sites internet dont la principale source de revenus est issue de la publicité pour la plupart d'entre eux, sans contribuer à leur maintien et en faisant reposer leur survie entièrement sur les autres internautes qui consultent ces sites. Il est dans ce cas possible de pousser le préjudice encore plus loin, dans le sens où le fait qu'une partie de la population qui fréquente un site internet utilise un bloqueur de publicité va nécessairement faire baisser les revenus liés à la publicité, et incitera donc les propriétaires du site à augmenter le nombre d'encarts publicitaires afin de restaurer leur niveau de revenus. Les premières victimes de cette augmentation de la quantité de publicités seront les internautes qui ne possèdent pas de bloqueur de publicités. Ce phénomène illustre l'effet Matthieu, à savoir que les pauvres (ici ceux qui n'utilisent pas de bloqueur de publicités) deviennent plus pauvres. Nous pointons dans le même temps ici un autre problème lié à l'obfuscation, à savoir qu'elle n'est rendue possible que parce qu'une minorité la pratique. En effet, par exemple si tous les internautes possédaient Adblock Plus, les sites internet dont le modèle économique repose sur la publicité verraient leurs sources de revenus s'effondrer et ne pourraient donc pas continuer leur activité.

Il est également reproché à l'obfuscation d'entacher le bon fonctionnement de l'internet, car il introduit du bruit dans le vaste canal de communication qu'il représente en y transmettant des données fausses ou altérées. Ces pratiques consomment donc de la bande passante et des capacités de stockage inutilement, et même à valeur ajoutée négatives puisque la transmission de ces fausses données empêche le bon fonctionnement d'un certain nombre de services. Cette critique n'est néanmoins pas la plus fréquemment émise contre l'obfuscation car les quantités de stockage et la bande passante disponibles ne constituent pas une contrainte actuellement, leur coût étant relativement faible. La tendance est d'ailleurs plutôt à stocker toutes les données et traces captables, y compris celles qui n'ont pas d'utilité immédiate. Néanmoins, là où cette introduction de bruit constitue un problème moral, selon nous et comme le font remarquer Finn Brunton et Helen Nissenbaum, c'est quand elle introduit des marges d'erreurs dans des usages du Big Data bénéfiques pour l'humanité, comme par exemple pour des expériences scientifiques, des études pour faire évoluer la médecine, … Nous en revenons ici au problème énoncé précédemment, et selon lequel l'obfuscation ne peut être viable que si elle est pratiquée par une minorité d'individus : si peu la pratiquent, l'erreur introduite dans les résultats des études et recherches sera négligeable, mais dès lors qu'ils deviennent nombreux, les résultats seront faussés et inutilisables, voire même risqueront d'induire en erreur les chercheurs dans leurs conclusions. Les nouvelles possibilités

de captation de données puis d'analyse et de corrélation apportées par le Big Data à la science et dont nous avons dressé un aperçu qui se veut relativement exhaustif précédemment, deviendraient alors inutiles voire néfastes.

Enfin, le problème le plus souvent soulevé concernant les pratiques d'obfuscation, au-delà des questions éthiques qu'elles posent, est que le fait de falsifier nos informations afin de tromper les entités de collecte des données peut-être dangereux, car ces fausses données peuvent s'avérer correspondre à celles d'un autre individu. Ces pratiquent peuvent donc avoir des implications directes dans la vie des individus. Il extrêmement désagréable, en plus du fait que cela porte atteinte à la vie privée, de voir les informations nous concernant utilisées par d'autres. Mais cela pourrait surtout devenir extrêmement problématique dans un avenir proche du fait de la généralisation de l'utilisation des algorithmes du Big Data dans le domaine de la sécurité. Par exemple, dans le cadre de l'utilisation des systèmes de profilage par la police, des erreurs sur les personnes pourraient survenir du fait des pratiques d'obfuscation. Si un individu effectue des activités illégales sur internet, et que via les méthodes d'obfuscation il s'est malencontreusement retrouvé à utiliser l'identité d'un autre, cet autre pourrait alors être accusé à tort d'un crime qu'il n'a pas commis, et il lui serait extrêmement difficile de prouver sa bonne foi. Cet argument est toutefois à nuancer, comme nous le font remarquer Finn Brunton et Helen Nissenbaum, car la propagation d'informations incorrectes sur le web pourrait rendre les erreurs de ces systèmes extrêmement fréquentes et par-là rendre leur utilisation inacceptable, car trop peu fiable. Ce type de systèmes n'est à priori utilisé que dans le cas où le taux d'erreurs est très faible, du fait de l'extrême sensibilité des sujets dont ils traitent. Cette position peut être éthiquement défendable dès lors qu'on considère ces systèmes comme indésirables, ou du moins moralement discutables. Cette question est néanmoins sujette à controverse, surtout concernant les usages relatifs à la sécurité, car ce sujet est sensible et beaucoup estiment que leur sécurité doit être privilégiée au respect de leur vie privée, en est témoin la loi sur le renseignement en cours de discussion en France.

On constate qu'il existe de nombreux points de débat concernant la validité éthique de la collecte massive des données pour alimenter le Big Data d'une part, et des pratiques permettant d'éviter ou de tromper cette collecte d'autre part. Afin de dépasser cette impasse, Finn Brunton et Helen Nissenbaum proposent de considérer la fin visée pour juger de la validité morale de l'obfuscation, plutôt que la pratique elle-même. A savoir que si la pratique d'obfuscation est destinée à se protéger d'un usage excessif ou malintentionné du Big Data, celle-ci peut être jugée comme moralement acceptable, même si les moyens qu'elle emploie sont discutables, ou du moins questionnables. Pour reprendre l'exemple de Kant, il serait acceptable de mentir pour éviter qu'un crime ne soit commis, car le préjudice porté par le mensonge est moins important que celui porté à la vie d'autrui. Les deux enseignants nous donnent l'exemple de l'utilisation de l'obfuscation par des bandits pour braquer une banque, auquel cas la fin visée, et donc la pratique de l'obfuscation, sont discutables ; mais dans le cas où une banque fait usage de l'obfuscation pour protéger les données bancaires de ses clients, celle-ci est acceptable. Comme ils le relèvent, nous faisons là aussi face à un nouveau problème puisque les fins revendiquées par ceux qui pratiquent l'obfuscation font très souvent

débat. On leur reproche par exemple souvent d'en faire usage pour cacher des téléchargement illégaux alors qu'ils revendiquent simplement vouloir échanger anonymement des données dont ils sont propriétaires. Toutefois, les fins visées par les pratiques d'obfuscation sont au moins partiellement légitimes, dans le sens où elles questionnent justement la légitimité des pratiques de collecte et de fouille de données ; or cette légitimité mérite, comme nous l'avons montré précédemment, d'être discutée. Il n'en reste pas moins que faire usage de l'obfuscation revient à accepter et revendiquer l'idée selon laquelle il y a des problèmes de violation des données personnelles dans les pratiques actuelles du Big Data, tout en refusant de se priver des services qu'il rend, et en acceptant le fait que le coût repose sur ceux qui ne se protègent pas, par volonté de se prêter au jeu ou, trop souvent, par ignorance. Les obfuscateurs font malgré tout face à un problème éthique.

Conclusion

Nous avons donc montré au cours de cet exposé que le Big Data influe désormais sur pratiquement tous les aspects de la vie des individus : à savoir leurs comportements individuels comme la façon dont ils accèdent à l'information, communiquent entre eux, consomment et se laissent tenter par les biens de consommation, et même jusqu'à l'image qu'ils se font d'eux-mêmes. Mais aussi sur notre façon de pratiquer et concevoir la science et la recherche, avec notamment les possibilités nouvelles offertes par le Big Data pour la recherche médicale. Enfin, les modes de gouvernement de nos sociétés s'en trouvent également bouleversés, avec la tendance notamment vers un monde dans lequel les comportements dits déviants sont rendus impossibles, plutôt que vers une volonté de dissuader les individus de pratiquer ces comportements, comme c'était le cas jusqu'alors. Le Big Data apporte ainsi de nouvelles problématiques en termes de sécurité, de législation et de justice. Nous avons dans un deuxième temps pointé le fait que la légitimité du Big Data repose sur une acceptation du public des algorithmes sur lesquels il repose, notamment en vertu d'une prétendue objectivité qui leur est conférée. Ce qui nous a permis de pointer les limites de cette réputation et de remettre en question cette notion d'objectivité, et ce notamment pour la raison qu'un algorithme n'est jamais que la description d'une suite d'étapes qui offrent un résultat en vue d'un objectif donné. L'algorithme ne livre donc que ce qu'on lui a demandé de livrer, il n'a pas d'intention, sa seule intention étant celle que lui a donnée celui qui l'a écrit. L'algorithme est donc le reflet de l'intention d'un ou plusieurs individus, or la notion d'objectivité parait difficilement applicable à l'individu, qui sert toujours un but. Dans un troisième et dernier temps, partant du constat que l'intrusion du Big Data dans la vie personnelle et intime des individus est extrêmement poussée, et par-là certainement excessive, nous avons tâché de donner un aperçu des alternatives qui s'offrent à nous pour nous en prévenir, et pointé les problèmes éthiques qu'impliquent ces alternatives, notamment parce qu'elles font trop souvent reposer le poids de ces intrusions sur les autres individus, qui ne s'en préviennent pas.

Le présent exposé a peut être laissé chez le lecteur l'impression que nous tendons vers un monde où tout sera contrôlé, et tous les individus surveillés et traqués. Cet état de fait représente bien évidemment un cas extrême, et nous avons bon espoir que l'humain saura

s'approprier cet outil qu'est le Big Data, et qui offre indéniablement d'infinies possibilités d'avancées positives. Nous reprendrons l'idée de Wendy Hui Kyong, reprise par John Cheney-Lippold : « *internet is neither a tool for freedom nor a tool for total control. Control is never complete, and neither is our freedom* » [Cheney-Lippold, 2011, p177]. Internet ne serait donc que le reflet de ce que nous en faisons, et s'il n'y a certes pas de « *contrôle total* » [Cheney-Lippold, 2011], nous admettrons toutefois que les technologies du Big Data vont rendre possible de s'en approcher à un point incomparable à ceux jamais atteints jusqu'alors car il multiplie les possibilités de contrôle : notamment parce que, comme le souligne John Cheney-Lippold, les entreprises du Big Data ont rendu pratiquement impossible de ne pas être surveillés lorsque nous naviguons sur internet. Cela est principalement dû au fait que, comme le rapporte Louise Merzeau, sur internet, contrairement à la lecture d'un journal ou de regarder la télévision, l'individu est obligé de laisser des traces, il ne peut pas rester passif, il doit naviguer entre les plateformes auxquelles il accède. Elle explicite d'ailleurs très clairement cette idée : « *De la même façon qu'on ne peut pas ne pas communiquer, on ne peut pas ne pas laisser de traces* » [Merzeau, 2013, p8]

Il nous semble également important d'aborder un autre point fort du problème que nous avons tâché de traiter : cet exposé a mis en valeur un basculement du pouvoir. Partant de l'hypothèse que la connaissance est la condition nécessaire et suffisante pour l'accès au pouvoir, on peut admettre qu'il est aujourd'hui détenu par ceux qui conçoivent et construisent le Big Data, à savoir les ingénieurs et les scientifiques. Alors qu'autrefois le savoir était entre les mains des « *sachants* », ceux qui possédaient une culture large, il est désormais entre les mains de ceux qui ont la capacité de la diffuser. Aldous Huxley, écrivain britannique, résume assez bien notre pensée à ce sujet :

« *Savoir c'est pouvoir, et par un paradoxe apparent il se trouve maintenant que ce sont les scientifiques et les techniciens qui, grâce à leur savoir de ce qui se passe dans un monde non-vécu d'abstractions et de déductions, ont acquis cette puissance immense et croissante qui est la leur, dirigent et modifient le monde dans lequel les hommes ont à la fois le privilège et l'obligation de vivre.* » [Huxley]

Cette réflexion sur l'idée de savoir nous amène à nous poser une autre question, liée à l'explosion de la quantité d'informations disponibles grâce à internet. Nous ne laissons plus de place à l'oubli, la moindre parcelle d'information étant stockée par défaut, et les individus ont accès à la très large majorité de la quantité d'informations disponibles mondialement en permanence et depuis n'importe quel endroit. Pourtant, notre cerveau est construit de telle façon qu'il sélectionne parmi les informations qu'il capte et choisit d'en conserver certaines et d'en oublier d'autres. Cette capacité d'oubli permet aux individus de se construire et d'avancer. Comment aller de l'avant dès lors que nous avons continuellement en tête une humiliation vécue plus tôt ? L'un des enjeux du Big Data sera donc, à notre sens, de savoir quel rapport à l'information nous souhaitons avoir, et de décider ce qu'il faut retenir, ce qu'il faut oublier, car ce ne sera bientôt plus notre cerveau qui en décidera, mais les objets numériques qui nous entoureront.

Bibliographie

- Crawford K. & Schultz J., (2014), *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*
- Weisberg J., (2011), Bubble trouble: *is web personalization turning us into solipsistic twits ?*, Slate Magazine
- Morozov E., (2011), *Your own Facts*, New York Times
- Eynard J., (2012), *L'éthique à l'épreuve des nouvelles particularités et fonctions des informations personnelles*, Ethique Publique, vol. 14, n° 2
- Gillepsie T., (2013), *The relevance of algorithms*, Media Technologies
- Anderson C., (2008), *The end of theory : The data deluge makes the scientific method obsolete*, Wired, Magazine 16.07
- Foucault M., (1978), *La « gouvernementalité* », Dits Ecrits, Tome III, n° 239
- Cheney-Lippold J., (2011), *A new algorithmic identity : soft biopolitics and the modulation of control*, Theory Culture Society, 28:164
- Brunton F & Nissenbaum H., (2011), *Vernacular resistance to data collection and analysis : A political theory of obfuscation*
- Cardon D., (2013), *Dans l'esprit du PageRank, Une enquête sur l'algorithme de Google*, Cairn, La Découverte, Réseaux n°177, 63:95
- Conseil d'Etat, (2014), *Le numérique et les droits fondamentaux*
- Grimmelmann J., (2008), *The Google dilemma,* Volume 53
- Merzeau L., (2013), *L'intelligence des traces*, Intellectica, n°13
- Morozov, *Your own facts*,
- Rouvroy A. & Berns T., (2013), *Gouvernementalité algorithmique et perspectives d'émancipation : Le disparate comme condition d'individuation par la relation ?*, Cairn, La Découverte, Réseaux n°177, 163 :196
- Rouvroy A. & Berns T., (2013), *Le nouveau pouvoir statistique, ou quand le contrôle s'exerce sur un réel normé, docile et sans événement car constitué de corps « numériques »*, Cairn, Association Multitudes, Multitudes, n°40, 163 :196
- Weisberg J., (2011), *Bubble trouble : Is web personalization turning us into solipsistic twits ?*, www.slate.com[29]

.

[29] Article de Jacob Weisberg disponible à cette adresse : http://www.slate.com/articles/news_and_politics/the_big_idea/2011/06/bubble_trouble.html