Dominique Cardon

À quoi rêvent les algorithmes

Nos vies à l'heure des big data

LA REPUBLIQUE DES IDEES

Seuil

Dominique Cardon

À quoi rêvent les algorithmes

Nos vies à l'heure des *big data*

LA REPUBLIQUE DES IDEES | Seuil

Collection dirigée
par Pierre Rosanvallon
et Ivan Jablonka

ISBN : 978-2-02-127998-6

www.seuil.com

INTRODUCTION

Comprendre la révolution des calculs

Un nouvel objet a fait son entrée dans nos vies : les algorithmes.

Ce terme d'informatique a une signification bien plus large qu'on ne le croit. Comme la recette de cuisine, un algorithme est une série d'instructions permettant d'obtenir un résultat. À très grande vitesse, il opère un ensemble de calculs à partir de gigantesques masses de données (les « *big data* »). Il hiérarchise l'information, devine ce qui nous intéresse, sélectionne les biens que nous préférons et s'efforce de nous suppléer dans de nombreuses tâches. Nous fabriquons ces calculateurs, mais en retour ils nous construisent.

Il n'est plus beaucoup de gestes quotidiens, d'achats, de déplacements, de décisions personnelles ou professionnelles qui ne soient orientés par une infrastructure de calculs. Quand elle vient soudainement à disparaître, comme lorsqu'une panne interrompt le trafic téléphonique, nous sommes désemparés. Pourtant, dès que nous pensons à la présence des calculateurs dans nos sociétés, nous maudissons la froide rationalité des machines et redoutons qu'elles ne prennent le pouvoir sur nous. Nous aimons leur opposer « notre » subtile sagacité.

Pourtant, les technologies trament notre monde depuis si longtemps qu'il est erroné de séparer les humains de leur environnement sociotechnique. Des premiers outils préhistoriques à l'invention de l'écriture, de la mécanisation de l'imprimerie à la numérisation de l'information, de la création des listes et des tableaux comptables au calcul scientifique, la longue histoire des technologies intellectuelles est au cœur de l'évolution de l'humanité. Il serait naïf de croire qu'elles n'ont pas transformé

profondément ce que nous sommes, ce que nous savons, nos manières de penser et les représentations que nous avons de nous-mêmes. Nous vivons dans une telle proximité avec les technologies que ce couple ne peut plus être défait sans que nous amputions la meilleure part de nous-mêmes.

Comme l'invention du microscope a ouvert une nouvelle fenêtre sur la nature, les capteurs numériques sont en train de jeter leur filet sur le monde pour le rendre *mesurable en tout*. Le savoir et les connaissances, les photographies et les vidéos, nos mails et ce que nous racontons sur Internet, mais aussi nos clics, nos conversations, nos achats, notre corps, nos finances ou notre sommeil deviennent des données calculables.

Aussi est-il essentiel de comprendre, de discuter et de critiquer la manière dont les algorithmes impriment leurs marques sur nos existences, jusqu'à devenir indiscutables et même invisibles. L'objet de ce livre est de comprendre ce que la révolution des calculs apportée par les *big data* est en train de faire à nos sociétés. Il décrit le monde auquel rêvent les algorithmes, avant que nous nous réveillions – trop tard.

Chiffrer le monde

Avant leur spectaculaire entrée dans nos vies quotidiennes, les calculs étaient surtout l'affaire d'États et d'entreprises. Longtemps, la mesure statistique a été une question de spécialistes. Le grand public n'en percevait l'écho qu'à travers la publication d'indicateurs simplifiés venant justifier des choix de politiques publiques. Indispensable colonne vertébrale des États et des marchés, les grandes institutions statistiques ont très vite été gouvernées par des professionnels de la mesure, usant d'outils et de modèles de plus en plus complexes.

Instruments de connaissance, les statistiques étaient aussi conçues comme des instruments politiques aux mains des décideurs. En « photographiant » le monde, elles donnaient aux hommes de pouvoir des outils pour évaluer, choisir et faire agir[1]. Depuis leur tour d'ivoire, statisticiens, sociologues et économètres veillaient à ce que l'existence des mesures n'influence pas le comportement des « mesurés ». À partir des politiques néolibérales des années 1980, on assiste à une généralisation de la calculabilité et à une

systématisation de la politique des indicateurs. La présence des quantificateurs dans la vie sociale se fait partout sentir. Baromètres, indices et palmarès entreprennent de chiffrer des activités qui, jusqu'alors, n'étaient pas mesurées ou dont la quantification ne faisait pas l'objet d'une attention constante et inquiète.

Les instruments statistiques sont devenus une technique de gouvernement. L'évaluation des politiques publiques en fonction d'objectifs chiffrés s'est généralisée. Les palmarès d'écoles, d'hôpitaux ou de régions où il fait bon vivre font la une des magazines. Les outils de gestion s'introduisent dans les activités les plus quotidiennes des salariés. Les systèmes de notation financière branchent leurs résultats sur une interminable chaîne de mécanismes comptables[2]. Sous prétexte d'efficacité, les indicateurs se sont répandus dans la société pour fournir à ceux qui étaient mesurés des chiffres destinés à orienter leurs comportements[3].

L'objectif de ces indicateurs est moins de connaître le réel que de « conduire les conduites[4] » des individus pour qu'ils le transforment. Les statisticiens traditionnels se sont trouvés désemparés devant ce déluge de chiffres peu fiables, mais, désormais, ils n'ont guère de prise sur la manière dont les entreprises et les administrations se nourrissent, jusqu'à l'asphyxie, de chiffres destinés à comparer et à évaluer, dans une logique de compétition et de performance. Le tournant de la « politique des indicateurs », qui a vu les statistiques descendre dans le monde social, continue d'étendre les dispositifs de commensuration à un nombre toujours plus important de secteurs d'activités[5].

Aujourd'hui, une nouvelle vague d'extension de la calculabilité est en marche. Son ampleur est inédite et ses conséquences, bien qu'encore difficiles à évaluer, sont considérables. Sur la logique des indicateurs chiffrés se greffe désormais celle du calcul algorithmique embarqué à l'intérieur des interfaces numériques. En rencontrant l'informatique, les chiffres sont devenus des signaux numériques (listes, boutons, compteurs, recommandations, fils d'actualité, publicité personnalisée, trajet GPS, etc.) qui habillent toutes les interfaces que, d'un clic, nous ne cessons de caresser. Ils pénètrent si intimement notre vie quotidienne que nous percevons mal les longues chaînes qui conduisent des sympathiques écrans colorés aux

grandes infrastructures statistiques que la révolution numérique installe dans de lointains serveurs de données.

À très grande vitesse, un nombre croissant de domaines – la culture, le savoir et l'information, mais aussi la santé, la ville, les transports, le travail, la finance et même l'amour et le sexe – sont désormais outillés par des algorithmes. Ils organisent et structurent les informations, aident à prendre des décisions ou automatisent des processus que nous avions l'habitude de contrôler nous-mêmes. Deux dynamiques s'avancent pour nous faire entrer dans cette nouvelle *société des calculs*.

La première est l'accélération du processus de numérisation de nos sociétés, qui nourrit de gigantesques bases de données d'informations, lesquelles n'avaient jamais été enregistrées, rendues accessibles et facilement manipulables. Un torrent de données se déverse aujourd'hui sur Internet. Chaque jour, 3,3 milliards de requêtes sont effectuées sur les 30 000 milliards de pages indexées par Google ; plus de 350 millions de photos et 4,5 milliards de *likes* sont distribués sur Facebook ; 144 milliards d'e-mails sont échangés par 3 milliards d'internautes. Si l'on numérisait toutes les communications et les écrits depuis l'aube de l'humanité jusqu'en 2003, il faudrait 5 milliards de gigabits pour les mettre en mémoire. Aujourd'hui, nous générons ce volume d'informations numériques en deux jours !

À l'instar des grandes révolutions industrielles, toutes initiées par l'exploitation d'un nouveau type d'énergie, le « nouvel or » des données numériques constitue, pour les promoteurs des *big data*, un gisement de valeur susceptible de relancer l'innovation, la productivité et la croissance. Aussi invitent-ils les institutions et les entreprises à conduire des politiques d'ouverture des données (*open data*), les citoyens à s'emparer des données pour surveiller les puissants (*sousveillance*) et les médias à pratiquer le « journalisme de données » (*data journalism*).

Si le discours public se concentre aujourd'hui sur le volume extravagant des données numérisées et les menaces que leur extraction fait peser sur la vie privée des individus, le principal défi que doivent affronter les *big data* est de donner du sens à ce magma de données brutes. Aussi la deuxième dynamique qui nourrit la société des calculs est-elle le développement de procédés, les algorithmes, donnant aux ordinateurs des instructions

mathématiques pour trier, traiter, agréger et représenter les informations. Venues de mondes différents, ceux du marketing, des marchés financiers ou de l'actuariat, de puissantes techniques statistiques (notamment celles que l'on qualifie de « prédictives ») se déploient à grande échelle en profitant de l'exceptionnelle augmentation des capacités de calcul des ordinateurs.

Greffés à nos écrans, les classements, palmarès, compteurs, cartes, recommandations et notes de toutes sortes figurent les pointes émergées de la calculabilité des traces de nos activités. À partir de données toujours plus inattendues (déplacements, tickets de caisse, clics sur Internet, consommation électronique, temps de lecture d'un livre sur les tablettes électroniques, nombre de pas enregistrés par un podomètre), les algorithmes chiffrent le monde, le classent et prédisent notre avenir.

Ouvrir la boîte noire

Omniprésents, ces calculs restent pour nous mystérieux. Ils orientent des décisions, appareillent des processus automatiques et justifient des choix politiques, mais nous interrogeons rarement la manière dont ils ont été produits. Nous regardons leurs effets sans examiner leur fabrication. Quelles sont les données qui servent au calcul ? Comment l'information a-t-elle été quantifiée ? Quels sont les principes de représentation qui animent le modèle statistique mis en œuvre pour classer tel objet plutôt que tel autre ? Qui pilote le codage des calculs et quels sont ses objectifs ?

Habités par un sentiment d'incompétence, nous préférons ignorer les conditions de fonctionnement de la société des calculs, en laissant les clés aux statisticiens, aux informaticiens et aux économètres. La complexification des modèles algorithmiques mis en œuvre dans les nouvelles infrastructures informationnelles contribue à imposer le silence à ceux qui sont soumis à leurs effets. Elle désarme aussi ceux qui entreprennent de critiquer l'avènement de la froide rationalité des calculs, sans chercher à en comprendre le fonctionnement. Par facilité autant que par ignorance, la critique du nouvel empire des calculs s'est réfugiée dans une pseudo-opposition entre les « humains » et les « machines ». Elle dénonce

confortablement la rationalisation néolibérale du monde, la tyrannie de l'évaluation ou les accidents automatisés du *high-frequency trading*.

Si cette défiance constitue une sorte de contrepoison, elle reste plus gratifiante que véritablement efficace. La critique de la raison calculatoire ne peut opposer qu'une rêverie pastorale à la marche automatisée des grands systèmes technologiques mondiaux. Pour vraiment critiquer une dynamique qui possède de si puissants moteurs économiques et culturels, il est nécessaire d'*entrer dans les calculs*, d'explorer leurs rouages et d'identifier leurs visions du monde. Avant de réduire la logique calculatoire aux intérêts économiques de ceux qui la fabriquent, il faut commencer par allonger les algorithmes sur le divan et entendre la variété de leurs désirs. Cet examen est indispensable si l'on veut débattre publiquement des calculs que nous voulons et de ceux dont nous ne voulons pas, contrôler leurs agissements et leur opposer des calculs alternatifs. Une radiographie critique des algorithmes est un enjeu démocratique aussi essentiel qu'inaperçu.

L'objet de ce livre est d'éclairer les enjeux sociaux, éthiques et politiques qui accompagnent le développement du calcul algorithmique, en prêtant attention au principal foyer des bouleversements en cours : celui des données numériques et, plus spécifiquement, du classement de l'information sur le web. Ma conviction est que, face au déploiement de la société des calculs, il est nécessaire d'encourager la diffusion d'une culture statistique vers un public beaucoup plus large que celui des seuls spécialistes.

Mais le propos de ce livre n'est pas mathématique : il est pleinement politique. La manière dont nous fabriquons les outils de calculs, dont ils produisent des significations, dont nous utilisons leurs résultats, trame les mondes sociaux dans lesquels nous sommes amenés à vivre, à penser et à juger. Les calculs habitent nos sociétés bien plus centralement que ne l'imaginent ceux qui voudraient les réduire à des fonctions mathématiques et rejeter la technique hors de la société, comme un *alien* menaçant. Les calculateurs fabriquent notre réel, l'organisent et l'orientent. Ils produisent des conventions et des systèmes d'équivalence qui sélectionnent certains objets au détriment d'autres, imposent une hiérarchisation des valeurs qui en vient progressivement à *dessiner les cadres cognitifs et culturels* de nos sociétés.

Comme l'ont souligné beaucoup de travaux d'histoire et de sociologie, les objets techniques ne fonctionnent que parce qu'ils opèrent dans un « milieu associé » qui les rend efficaces et pertinents[6]. Les calculs ne calculent vraiment que dans une société qui a pris des plis spécifiques pour se rendre calculable. Aussi faut-il comprendre comment nos sociétés sécrètent certaines manières de se chiffrer plutôt que d'autres. Que valorisent-elles dans leur façon de compter et de classer ?

Il suffit d'ouvrir la boîte noire des calculateurs pour constater qu'ils servent des desseins très différents. Selon la nature des données enregistrées, la manière de les catégoriser, le choix des techniques statistiques ou les options de visualisation des résultats, le fait de modifier les paramètres du calcul conduit à valoriser des choses très différentes. Face aux visées productivistes de la mesure du PIB, des économistes hétérodoxes opposent d'autres « indicateurs de richesse », comme l'indice de développement humain (IDH) du Programme des Nations unies pour le développement (PNUD) popularisé par Amartya Sen. Ils voudraient déplacer le centre de gravité sur lequel reposent les calculs macro-économiques mondiaux vers la prise en compte de nouvelles variables, comme l'espérance de vie à la naissance, le niveau d'éducation, la qualité de vie ou le bonheur[7].

En modifiant la traditionnelle mesure de répartition des revenus par décile pour la décomposer en centiles, Camille Landais et Thomas Piketty ont fait apparaître l'explosion récente des écarts de richesse en faveur du 1 % de la population, qui n'apparaissait pas avec un filtre moins fin. La perspective nouvelle offerte par ce changement de lunette statistique a inspiré le slogan « Nous sommes les 99 % » au sein du mouvement des Indignés et Occupy, au début des années 2010[8]. Il ne faut pas grand-chose – choisir d'autres variables, changer l'échelle, calculer autrement – pour faire du chiffrage la meilleure arme contre d'autres chiffrages[9]. Il est crucial de lutter contre cette sorte de fatalisme qui nous conduit à imputer aux mesures ce que, en réalité, nous leur avons demandé de faire.

On croit volontiers qu'un unique moteur anime la guerre de conquête entreprise par les calculs : la performance économique. Dans cet ouvrage, on n'abordera pas directement les enjeux économiques de la domination des grandes plateformes du web, les fameux GAFA (Google, Apple, Facebook,

Amazon). Leurs ambitions, leurs intérêts, leur culture californienne sont à la une des magazines et sont désormais bien connus. Ce livre ne propose pas de critiquer les algorithmes de l'extérieur, en en faisant les reflets des intérêts de leurs concepteurs, mais de comprendre de l'intérieur la manière dont ils produisent des effets (plus ou moins critiquables) sur nos sociétés.

C'est, la plupart du temps, au nom de l'efficacité que les techniques calculatoires colonisent des univers toujours plus nombreux. Les recherches sur Google sont de plus en plus personnalisées, afin de répondre au mieux à nos attentes et d'anticiper des désirs que nous ne connaissons pas encore. Amazon voudrait nous envoyer des livres avant même que nous ne les ayons commandés tellement, forte de ses calculs, l'entreprise pense savoir ce que nous voudrions lire. Depuis quelques années, le marché de la publicité numérique nourrit l'espoir que, devenue « personnelle », la publicité ciblée perdra son caractère intempestif pour devenir, aux yeux de ceux qu'elle vise, une information comme une autre.

Ces manières de chiffrer l'information font souvent l'objet de critiques. Elles enferment les individus dans la bulle de leurs propres choix, plient leur destin dans l'entonnoir du probable et nourrissent la précision du ciblage d'une capture disproportionnée d'informations personnelles. Mais elles n'adviennent que parce qu'elles font écho à des transformations des modes de vie et des aspirations que suscitent les processus d'individualisation de nos sociétés. La thèse de ce livre est que, si les logiques de personnalisation s'installent aujourd'hui dans nos vies, c'est parce qu'elles calculent une forme nouvelle du social, la société des comportements, où se recompose la relation entre le centre de la société et des individus de plus en plus autonomes.

Notes

1. Alain Desrosières, *La Politique des grands nombres. Histoire de la raison statistique*, Paris, La Découverte, 2000.

2. Pierre Lascoumes et Patrick Le Galès (dir.), *Gouverner par les instruments*, Paris, Les Presses de Sciences Po, 2004.

3. Isabelle Bruno et Emmanuel Didier, *Benchmarking. L'État sous pression statistique*, Paris, Zones, 2013.

4. Michel Foucault, « Le sujet et le pouvoir », in *Dits et écrits*, Paris, Gallimard, coll. « Quarto », 1982, tome II, p. 1041-1062.

5. Barbara Cassin (dir.), *Derrière les grilles. Sortons du tout-évaluation*, Paris, Mille et une nuits, 2014.

6. Gilbert Simondon, *Du mode d'existence des objets techniques*, Paris, Aubier, 1989 ; et Bruno Latour, *La Science en action*, Paris, La Découverte, 1989.

7. Éloi Laurent et Jacques Le Cacheux, *Un nouveau monde économique. Mesurer le bien-être et la soutenabilité au XXIe siècle*, Paris, Odile Jacob, 2015.

8. Camille Landais, « Les hauts revenus en France (1998-2006). Une explosion des inégalités ? », École d'économie de Paris [disponible sur *http://ur1.ca/ge0sw*].

9. Isabelle Bruno, Emmanuel Didier et Julien Prévieux (dir.), *Stat-Activisme*, *Comment lutter avec des nombres ?*, Paris, Zones, 2014.

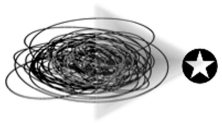
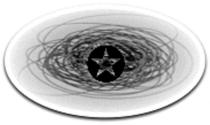
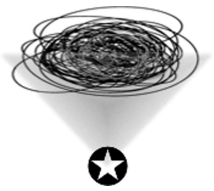
CHAPITRE PREMIER

Quatre familles de calcul numérique

Afin d'approcher de façon simple les enjeux qui président aux classements de l'information numérique, je proposerai un récit faisant se succéder quatre manières de produire de la visibilité avec des calculs. De façon métaphorique, on identifiera ces quatre familles en fonction de la place qu'occupe le calculateur par rapport au monde qu'il entend décrire. Les mesures peuvent se trouver *à côté*, *au-dessus*, *dans* ou *en dessous* des données numériques. La réalité des calculateurs du web n'a rien d'aussi géométrique. Mais le défi que voudrait relever cet ouvrage est de proposer une clé de lecture, afin que chacun se sente autorisé à ouvrir la boîte noire des algorithmes du web.

Résumons tout de suite, avant de les explorer, ces quatre familles. Les mesures d'audience, en premier lieu, se placent à côté du web pour dénombrer les clics des internautes et ordonner la *popularité* des sites. La famille de mesures issue du PageRank, l'algorithme de classement de l'information du moteur de recherche de Google, voudrait se situer au-dessus du web, afin de hiérarchiser l'*autorité* des sites au moyen des liens hypertextes qu'ils s'échangent. Les mesures de *réputation*, qui se sont développées avec les réseaux sociaux d'Internet et les sites de notation, se positionnent à l'intérieur du web, afin de donner aux internautes des compteurs qui valorisent la *réputation* des personnes et des produits. Enfin, les mesures *prédictives* destinées à personnaliser les informations présentées à l'utilisateur déploient, sous le web, des méthodes statistiques d'apprentissage pour calculer les traces de navigation des internautes et leur prédire leur comportement à partir de celui des autres.

Le parcours que nous allons entreprendre à travers ces quatre manières de classer l'information numérique permettra de dégager les différentes valeurs qui nourrissent les choix que font les algorithmes : la popularité, l'autorité, la réputation et la prédiction. On montrera notamment comment ceux-ci prélèvent sur le web des données différentes (clics, liens, *likes* et traces) pour les soumettre à des opérations répondant à différentes conventions statistiques que résume le tableau ci-dessous.

	À côté	**Au-dessus**	**Dans**	**Au-dessous**
				
Exemples	Médiamétrie, Google Analytics, affichage publicitaire	PageRank de Google, Digg, Wikipédia	Nombre d'amis Facebook, Retweet de Twitter, notes et avis	Recommandation Amazon, publicité comportementale
Données	Vues	Liens	Likes	Traces
Population	Échantillon représentatif	Vote censitaire, communautés	Réseau social, affinitaire, déclaratif	Comportements individuels implicites
Forme du calcul	Vote	Classements méritocratiques	*Benchmark*	*Machine learning*
Principe	***Popularité***	***Autorité***	***Réputation***	***Prédiction***

En distinguant ces quatre familles, souvent confondues et mal comprises, on voudrait faire apparaître à quel point les techniques de calcul justifient, selon des principes différents, la manière dont elles produisent un ordre de tel type plutôt que de tel autre. Entrer dans les calculs, c'est comprendre que les enjeux techniques et industriels qui opposent les acteurs du web nourrissent aussi une compétition sur la bonne manière d'organiser la visibilité des informations.

Afin de présenter ces quatre familles du calcul numérique, je suivrai un fil chronologique. Même si, au sein des actuelles plateformes du web, les

quatre familles de calcul cohabitent et se mélangent les unes aux autres, elles sont apparues successivement dans l'histoire d'Internet. Leur chronologie fait apparaître les points de bascule qui nourrissent les tensions actuelles entre les différents écosystèmes du web.

À côté du web : l'imprécise popularité des clics

La première technique de calcul se place à côté du web pour mesurer l'audience des sites et leur popularité. Alors que les pionniers d'Internet n'y prêtaient guère attention, quand le réseau a pris son essor, au début des années 1990, certains webmestres ont cherché à connaître le nombre de visiteurs sur leur site. Ils ont fabriqué un indicateur dont le principe est directement inspiré des techniques de mesure d'audience des médias de masse : compter les clics des visiteurs.

Afin d'éviter de dénombrer plusieurs fois le même internaute, la notion de « visiteur unique » – vérifiée à travers l'adresse IP de l'ordinateur – est la principale unité de compte de la popularité des médias en ligne et, par simple équivalence, du tarif des publicités qu'ils accueillent. La mesure d'audience mime le vote démocratique : chaque internaute qui clique dispose d'une voix et d'une seule, et ceux qui dominent le classement sont ceux qui ont attiré l'attention du plus grand nombre. Comme l'a montré l'histoire de la quantification des publics de la presse, de la radio et de la télévision, la mesure d'audience trouve sa légitimité dans son étroite proximité avec le nombre démocratique[1].

Public et électorat font cause commune. Ils partagent une même idée de la représentativité statistique, fondée sur le décompte de voix unitaires et semblables au sein du périmètre de référence de la nation. Ils s'organisent autour d'une asymétrie entre un centre restreint d'émetteurs (l'espace politique, l'espace médiatique) et une population silencieuse de récepteurs (les électeurs, les téléspectateurs). Au centre de la scène, quelques médias se répartissent les suffrages d'individus dispersés que le programme rassemble, éduque et fait communier à travers le partage d'une même expérience. Les programmes « populaires » fédèrent le « grand public » en faisant naître

cette « communauté imaginée » qui participe à la formation des représentations collectives des citoyens[2].

Cependant, avec la dérégulation croissante du secteur des médias et le rôle de plus en plus important de la publicité, la mesure d'audience sert moins à fabriquer le « public » qu'à mesurer des « parts de marché ». Dans l'univers numérique où l'offre d'information est abondante et incontrôlée, l'audience a perdu toute relation avec l'idée d'une représentation politique des publics. La mesure de la fréquentation des sites intéresse principalement le marché publicitaire et la toute petite fraction des éditeurs de sites qui en vivent. Le paradoxe de l'audience web est que cette mesure obsède les sites qui dominent les classements de popularité, mais qu'elle est en réalité très imprécise et de plus en plus contestée par les nouvelles techniques de personnalisation qui mesurent les individus sans se préoccuper de représenter les publics.

L'audience des sites web se mesure de deux façons[3]. Sur le modèle des médias de masse, elle peut d'abord être *user centric* : une sonde installée dans l'ordinateur d'un panel représentatif de la population des internautes français enregistre leur navigation pour, ensuite, classer l'audience des sites les plus consultés. Médiamétrie/NetRatings compte ainsi 20 000 personnes dans son échantillon représentatif des internautes français et ComScore prétend tracer 2 millions d'internautes dans le monde entier. Tous les mois, les sites les plus visités sont classés par leur popularité et ce classement détermine la valeur des bannières publicitaires[4]. Déjà imparfaite dans le monde de la télévision, cette mesure révèle de redoutables imprécisions lorsqu'elle est appliquée au web.

Car cette méthode ne sait jamais très bien qui se trouve derrière l'ordinateur familial. Elle a du mal à suivre le panéliste sur les différents terminaux qu'il utilise. Elle suppose que les parcours multiples, rapides et enchevêtrés d'une navigation sur le web équivalent à une information lue, vue ou entendue dans les médias traditionnels. Enfin, l'offre de sites à visiter étant sans commune mesure avec le nombre de chaînes de télévision ou de journaux, cette mesure ne parvient réellement à classer qu'une infime fraction de sites extrêmement populaires et centraux[5]. Conçue pour

départager les compétiteurs dans un contexte où l'offre était rare, la mesure d'audience ne sert de convention que pour la petite élite au sommet du web.

Devant tant d'imperfections, Médiamétrie propose une solution qui hybride une mesure *user-centric* avec l'autre technique d'enregistrement de l'audience web que l'on appelle *site-centric*. Les webmestres peuvent en effet connaître le trafic de leur site à l'aide de différents outils de supervision, dont le plus célèbre est *Google Analytics*. Ils enregistrent l'adresse IP de la machine qui s'est connectée, le site depuis lequel l'internaute est arrivé sur leur page ou la durée de consultation. Seul l'éditeur du site connaît ces informations et, souvent, lorsqu'il les rend publiques, il est tenté de publier des chiffres avantageux ou de les gonfler au moyen de techniques logicielles faciles d'emploi. Si le site souhaite participer à une mesure mutualisée des audiences web, il peut accepter que des entreprises externes comme Médiamétrie, Xiti ou Weborama accèdent aux données de logs enregistrées par un marqueur (*tag*) déposé par le webmestre sur chacune des pages de son site.

En l'absence d'une véritable régulation du secteur – qui ne se met en place que progressivement –, il est très facile de manipuler ces mesures d'audience, et les chiffres diffèrent sensiblement selon les instituts. Les médias d'actualité ne cessent, par exemple, de créer des jeux-concours attractifs afin de gonfler leurs chiffres. Soumis aux pressions concurrentielles d'un marché publicitaire qui paie peu, les sites d'informations cherchent à attirer de l'audience à travers des contenus divertissants « attrape-clics » (pratique qualifiée de *clickbait*).

Lorsque les chiffres de navigation viennent du réseau, les serveurs donnent des chiffres massifs et exhaustifs, mais dénombrent des logs de visiteurs étrangers, des robots cliqueurs et beaucoup de « bruits » informatiques. Cette mesure « machinique » de la fréquentation ne révèle ni qui est derrière le terminal, ni si la page ouverte a été lue, ni quelles sont les propriétés sociodémographiques des visiteurs. Aussi les praticiens de l'audience vont-ils essayer d'enrichir ces informations grâce au *cookie*. Quand, en 1994, Lou Montulli, un ingénieur de Netscape, a inventé le *cookie*, il s'agissait simplement de créer un petit fichier informatique déposé dans le navigateur de l'internaute, afin de se souvenir de l'adresse Internet

de la machine lorsqu'il constitue un panier d'achat en naviguant sur des pages différentes du site[6].

Ce fichier « mouchard » discrètement déposé dans le navigateur de l'internaute va devenir le cheval de Troie des publicitaires et des grandes plateformes du web pour pénétrer l'intimité des internautes. Le *cookie* permet au site visité de reconnaître l'internaute qui se connecte, afin de faciliter sa navigation (par exemple, en se souvenant de ses mots de passe), mais aussi de recueillir des informations plus indiscrètes sur ses navigations passées et de constituer un profil plus riche de l'utilisateur en agrégeant des informations de fréquentation et des informations sociodémographiques.

Entre mesure *user* et *site centric* se forment ainsi deux types de connaissance de l'audience qui dessinent une polarité que l'on va retrouver tout au long de cet ouvrage. D'un côté, les professionnels du marketing traditionnel s'intéressent à la qualification de leurs publics, au moyen des variables que le marketing et la sociologie ont produites ensemble : la profession, le revenu, l'âge, le style de vie ou le lieu d'habitation. Ils connaissent bien les individus, mais peu leurs comportements. De l'autre côté se met en place une connaissance par profil qui, à l'inverse, enregistre bien les comportements, mais sans vraiment connaître les individus. Cette tension sera au cœur de la recomposition des formats publicitaires dans le monde des traces enregistrées *sous* le web par la quatrième famille de calcul numérique.

Si les mesures d'audience se perfectionnent, elles doivent se défendre face à des webmestres peu scrupuleux, des éditeurs élargissant artificiellement le périmètre de leur site et des robots cliqueurs. Il ne fait guère de doute que la médiocrité de la mesure d'audience a contribué à la chute des tarifs publicitaires sur Internet. Mais cette métrique présente, pour les annonceurs, un autre défaut. Comme l'affichage publicitaire classique, elle enregistre la fréquentation des sites et non l'efficacité des messages. Dans les environnements numériques, il est possible de savoir si l'individu exposé à un message a interagi avec lui en cliquant sur l'annonce ou, mieux, en achetant sur le site de l'annonceur. C'est dans cette direction que vont s'orienter les publicitaires pour substituer, à la représentation démographique des publics, la mesure personnalisée de l'efficacité des messages.

Les médias traditionnels ont rapidement été domestiqués par la domination exclusive de la mesure d'audience. L'originalité d'Internet est d'avoir inventé d'autres manières de mesurer l'information et de la distribuer au public. Appliquée aux connaissances, la popularité n'est en rien garante de qualité. Elle valorise de façon écrasante les choix conformistes, consensuels et populaires. L'audience calcule le rayonnement qu'exerce un petit nombre d'émetteurs sur un public passif. Lorsque le public devient « actif », il formule une demande et souhaite trouver l'information de la meilleure qualité possible. Parce que le web bouleverse l'asymétrie entre une (petite) offre qui propose sans laisser beaucoup de choix et une (large) demande qui reçoit sans vraiment choisir, les pionniers du web ont opposé un autre système de classement qui ne s'appuie pas sur la popularité de l'information, mais sur son autorité.

Au-dessus du web : l'autorité des méritants

Avec l'arrivée de Google en 1998 s'est déployée, à une très large échelle, une nouvelle méthode statistique pour détecter la qualité de l'information en plaçant le calculateur *au-dessus* du web, afin d'enregistrer les échanges entre internautes sans les influencer. Inédite dans l'histoire des médias, cette solution est d'une grande audace. Avant Google, les premiers moteurs de recherche (Lycos, Alta Vista) étaient lexicaux : ils classaient mieux les sites qui contenaient le plus de fois le mot-clé de la requête de l'utilisateur dans leurs pages.

Sergey Brin et Larry Page, les fondateurs de Google, vont opposer à ce procédé inefficace une tout autre stratégie : plutôt que de demander à l'algorithme de comprendre ce qui dit la page, ils vont proposer de mesurer la *force sociale de la page* dans la structure du web. L'architecture particulière du réseau Internet fait du web un tissu de textes se citant les uns les autres à travers des liens hypertextes. L'algorithme du moteur de recherche ordonne les informations en considérant qu'un site qui reçoit d'un autre un lien reçoit en même temps un témoignage de reconnaissance qui lui donne de l'autorité. Sur ce principe, il classe les sites à partir d'un vote censitaire au fondement méritocratique. Les sites les mieux classés sont

ceux qui ont reçu le plus de liens hypertextes venant de sites qui ont, eux-mêmes, reçu le plus de liens hypertextes des autres[7]. Dans son principe initial, le PageRank, l'algorithme qui a fait la fortune de Google, considère que les liens hypertextes enferment la reconnaissance d'une autorité : si le site A adresse un lien vers le site B, c'est qu'il lui accorde de l'importance. Qu'il dise du bien ou du mal de B n'est pas la question ; ce qui importe est le fait que A ait jugé nécessaire de citer B comme une référence, une source, une preuve, un exemple ou un contre-exemple. Le seul fait de citer enferme le signal dont le calculateur fait son miel.

La culture de la communauté participative du web des pionniers rompt avec l'impératif de représentativité que les médias traditionnels imposaient à la figuration de leur public. L'information la plus visible n'est pas la plus vue, mais celle que les internautes agissants ont choisi de reconnaître en lui adressant beaucoup de liens. Les lecteurs silencieux sont oubliés et le dénombrement des liens n'a rien du vote démocratique. Plus un site est cité par les autres, plus la reconnaissance qu'il adresse à d'autres a de poids dans le calcul d'autorité. Empruntée au système de valeurs de la communauté scientifique et notamment aux classements des revues scientifiques qui donnent plus de poids aux articles les plus cités par les autres, cette mesure de reconnaissance a spectaculairement prouvé qu'elle constituait l'une des meilleures approximations possible de la qualité des informations.

Alors que les journalistes filtrent l'information sur la base d'un jugement humain avant de la publier, les moteurs de recherche (ainsi que Google News) filtrent *a posteriori* une information déjà publiée sur la base des jugements humains émis par l'ensemble des internautes qui publient sur le web. Dans l'univers numérique, ce principe a pris le nom d'« intelligence collective » ou de « sagesse des foules ». Il mesure l'information à partir des évaluations que s'échangent, de façon auto-organisée, les internautes les plus actifs. De nombreuses autres métriques confèrent de la même façon de l'autorité à ceux qui ont été reconnus par la communauté, comme sur Wikipédia ou sur Digg, dans le monde du logiciel libre, mais aussi dans le classement des avatars des jeux en ligne multijoueur.

Ces plateformes déploient des procédures permettant d'identifier la qualité des documents ou des personnes indépendamment de leur statut social, en mesurant l'autorité qu'ils ont acquise dans le réseau à travers les

jugements des autres internautes. Par approximations et révisions successives, et souvent avec un grand raffinement procédural, ces calculs ont pour ambition de rendre commensurables la vraisemblance, la pertinence et la justesse des informations avec, pour horizon, une conception exigeante de la rationalité et du savoir.

Une des particularités des mesures d'autorité est que le signal qu'elles enregistrent en se plaçant *au-dessus* du web ne doit pas être « contaminé » par les stratégies des internautes. Le rêve du PageRank de Google est que les internautes oublient son existence. La qualité de la mesure dépend étroitement du fait que ceux qu'elle mesure n'agissent pas en fonction de son existence. Ils doivent s'échanger des liens de façon « naturelle » et « authentique ». Pourtant, ce rêve naïf est constamment mis à mal par les stratégies de tous ceux qui cherchent à obtenir de la visibilité sur le web. Le florissant marché du SEO (*Search Engine Optimization*) est constitué d'entreprises qui proposent aux sites d'améliorer leur classement dans les résultats de Google. Certains proposent de parfaire le design et l'écriture du site pour que les robots du moteur de recherche le comprennent mieux. Mais beaucoup d'autres tentent de produire une autorité artificielle.

À la manière de la claque théâtrale, les stratèges du marché du référencement paient ou fabriquent des sites qui citent leurs clients : ils placent des liens vers le site cible dans les commentaires de blogs, glissent subrepticement un lien dans Wikipédia, créent des « fermes » de faux sites liés les uns aux autres pour adresser ensuite un lien hypertexte vers la cible, produisent de faux contenus éditoriaux (parfois écrits par des robots) pour tromper l'algorithme. La plupart de ces techniques sont aujourd'hui devenues inefficaces en raison des modifications incessantes que Google apporte à l'algorithme pour décourager ceux qui essaient de tromper son classement. Mais ce jeu du chat et de la souris entre les webmestres et les concepteurs de l'algorithme est sans fin.

Si les mesures d'autorité affichent fièrement l'originalité du classement de l'information numérique par rapport aux médias traditionnels, elles ont aussi fait l'objet de nombreuses critiques dont la pertinence n'a cessé de se renforcer avec la massification des usages du réseau. Deux reproches vont conduire à un second tournant dans l'histoire des classements numériques.

Une première critique souligne que l'agrégation du jugement des pairs produit de puissants effets d'exclusion et de centralisation de l'autorité. Comme toute forme en réseau, ceux qui se trouvent au centre attirent l'attention de tous et reçoivent une visibilité imméritée. À force d'être cités par tous, les plus reconnus deviennent aussi les plus populaires et reçoivent en conséquence le plus de clics[8]. L'aristocratique mesure d'autorité s'abîme alors en une vulgaire mesure de popularité. Elle est devenue, notamment pour Google, un attracteur de trafic permettant de valoriser la publicité achetée par les annonceurs que l'entreprise de Mountain View classe séparément dans la colonne des « liens sponsorisés ». Bien que distincts, le classement « naturel » par l'autorité et le classement « marchand » de la publicité font de la première page des résultats de recherche le grand carrefour du trafic sur le web, imposant soit les sites les plus centraux et conventionnels, soit les sites de ceux qui ont accepté de payer pour être vus à leurs côtés.

La seconde critique porte sur l'effet censitaire des mesures d'autorité. Ne participent au classement de l'information que ceux qui publient des documents comportant des liens hypertextes, comme les détenteurs de sites ou les blogueurs ; les autres sont ignorés. Or, avec la massification des usages d'Internet, s'inventent sur les réseaux sociaux d'autres manières de participer, plus volatiles et conversationnelles, plus spontanées et moins sélectives socialement. Les internautes lecteurs sont devenus acteurs du web à travers leurs pages Facebook ou leur compte Twitter. Les réseaux sociaux ont attiré vers eux des publics plus juvéniles, moins diplômés et plus divers socialement et géographiquement. La voix de ces nouveaux internautes actifs peut difficilement être ignorée des classements.

L'algorithme de Google fait « comme si » les liens hypertextes transportaient de la reconnaissance dont l'agrégation centrale produit de l'autorité. En revanche, il n'est plus possible de faire de même avec les *likes* de Facebook et les liens partagés sur Twitter. Ceux-ci sont investis de tant de significations subjectives, de jeux identitaires, d'appréciations contradictoires et d'idiosyncrasies contextuelles qu'un calculateur ne peut en extraire qu'un ordre très imparfait. Alors que le lien hypertexte projette une signification « calculable » vers l'ensemble du web, le *like* n'envoie de signification que vers le réseau social de l'émetteur.

Aussi les réseaux sociaux d'Internet proposent-ils d'éclater les classements, afin de les réorganiser par affinités autour du cercle d'« amis » ou de *followers* que s'est choisi l'internaute. Alors que les métriques d'autorité mesurent la reconnaissance dont les documents font l'objet indépendamment des qualités de leurs auteurs, c'est désormais la réputation numérique des internautes qui, autant que les documents eux-mêmes, va être évaluée.

À l'intérieur du web : la fabrique de la réputation

Alors que Google cherche à cacher le calculateur *au-dessus* du web afin que les internautes ne s'en emparent pas, les métriques de réputation du web social le glissent *dans* le web pour que les internautes se mesurent eux-mêmes. Le symbole de ces nouveaux calculs est le *like* de Facebook, pointe avancée d'un ensemble beaucoup plus large et disparate d'indicateurs mesurant la taille des réseaux personnels par le nombre d'amis, la réputation acquise en fonction du nombre d'informations publiées que d'autres internautes ont ensuite commentées ou partagées, le nombre de fois où le nom de l'internaute a été prononcé dans la conversation des autres, etc.

Les métriques de réputation mesurent le pouvoir qu'a l'internaute de voir les autres relayer les messages qu'il émet sur le réseau. L'influence procède toujours d'un ratio entre le nombre de personnes que l'on connaît et le nombre de personnes dont on est connu. Elle mesure la force sociale d'un nom, d'un portrait ou d'une image. La concurrence pour faire reconnaître ses arguments est devenue compétition pour assurer sa propre visibilité dans l'espace numérique. Le web social de Facebook, Twitter, Pinterest, Instagram, etc., s'est ainsi couvert de chiffres et de petits compteurs, des « gloriomètres », pour reprendre une expression de Gabriel Tarde. Ils dessinent un paysage hérissé de monticules et de vallées creuses, une topologie signalant les réputés, les influents et les notoires à ceux qui traversent la carte en utilisant les reliefs pour s'orienter.

Dans un univers où les compteurs sont omniprésents, rien n'interdit aux acteurs d'agir pour les classements. Alors que dans le monde de l'autorité,

la visibilité se mérite, dans celui des affinités numériques, elle peut se fabriquer. Façonner sa réputation, animer sa communauté d'admirateurs ou anticiper la viralité de ses messages constitue même un savoir-faire valorisé. S'il veut être retweeté, un tweet ne devra pas être envoyé le vendredi en fin d'après-midi, mais le lundi à 11 heures du matin, conseillent les manuels de management de la « e-réputation ». Placés dans le web, sous les yeux de tous, les compteurs rendent les internautes calculateurs. La métrique n'objective pas ; elle produit des signaux qu'utilisent les internautes pour orienter leur comportement et améliorer les scores qu'enregistrent les métriques.

Parallèlement aux mesures du web social, un autre ensemble de métriques de réputation est apparu sur le web, avec la mise en place du dispositif « Notes et avis » sur la plupart des sites de e-commerces. Hôtels, restaurants, produits culturels, et bientôt tout ce qu'il est possible d'évaluer et de noter, mettent désormais les internautes à contribution pour agréger leurs évaluations dans une opinion collective. C'est même une démocratisation du marché qui est figurée par l'idée que l'évaluation de la qualité par les consommateurs est susceptible de défaire l'asymétrie d'informations entre vendeurs et acheteurs et de se substituer aux experts jugés partiaux, autocentrés, voire corrompus.

Certains secteurs des services, comme l'hôtellerie et la restauration, ont été transformés par l'apparition de ces métriques de réputation. Dans le cas des biens culturels, comme l'évaluation des films cinématographiques, les notes ont plus d'importance qu'une collection d'avis singuliers[9]. Lorsque l'évaluation du bien comporte des aspects techniques, la parole d'experts est préférée à l'agrégation des notes de consommateurs peu compétents. La démocratisation de l'évaluation profane associe l'idée de pouvoir tout noter à celle de faire noter tout le monde.

Cependant, la participation à ces votes de paille non échantillonnés et volontaires engendre des effets complexes. Les notes et avis se concentrent, en réalité, sur une petite partie des produits ; une minorité active fabrique la majorité des évaluations ; les notes sont indulgentes et peu discriminantes. Aussi, pour stabiliser une mesure souvent biaisée par de nombreux faux avis, les plateformes doivent mettre en œuvre des dispositifs pour recalibrer

les notations en faisant évaluer non seulement les produits, mais aussi les commentateurs et en distinguant parmi eux une élite de connaisseurs[10].

Aussi originales soient-elles, les mesures de réputation ont à leur tour fait l'objet de critiques de plus en plus vives, avec la rapide diffusion des pratiques du web social. La première tient au fait que, choisissant d'éclater la visibilité en une myriade de compteurs pour lutter contre la centralité de la mesure d'autorité, ceux-ci sont accusés d'enfermer les utilisateurs dans une bulle. En choisissant leurs amis, les internautes font des choix affinitaires homogènes. Ils rassemblent des personnes dont les goûts, les centres d'intérêt et les opinions se ressemblent. En conséquence, les métriques basées sur les affinités numériques délimitent, pour l'utilisateur, des fenêtres de visibilité qui ont la couleur de leur réseau social, mais risquent du coup de faire disparaître des informations qui pourraient les surprendre, les déranger ou contredire leurs *a priori*[11].

La deuxième critique est que ces multiples mesures locales, en raison de leur hétérogénéité, peuvent difficilement être agrégées. Si certaines plateformes de « notes et avis » sont parvenues à développer auprès de leur public une véritable culture de l'évaluation, il n'existe pas en revanche de convention partagée pour représenter l'effervescence qui voit les internautes s'échanger des « demandes d'ami », des *likes* ou des *retweets*. La signification qu'enferment les micro-appréciations de réputation du web social est souvent trop jouée, trop calculée et, surtout, bien trop contextuelle pour être véritablement commensurable. La grande scène expressive des réseaux sociaux d'Internet met en lumière des signes, des désirs et des parades identitaires qui répondent à une économie de la reconnaissance et de la réputation. Sincères dans leur désir expressif, ces signes, agrégés et sortis de leur contexte, ne sont souvent ni véridiques ni authentiques.

Dans un espace où la visibilité est pensée comme l'effet d'une stratégie, un décalage de plus en plus important se creuse entre ce que les individus disent faire et ce qu'ils font réellement. Les études font apparaître une différence considérable entre le nombre des personnes qui déclarent regarder Arte et l'audience réelle d'Arte. Mais les réseaux sociaux ont installé sur le web une immense usine de production de signes expressifs, qui creuse un écart entre la multiplicité des désirs d'être et la réalité des

existences quotidiennes[12]. Cet écart déroute les projets des calculateurs pour donner du sens au grand bazar des données numériques. Ils ne savent plus s'ils doivent interpréter ce que disent les internautes – ce qu'ils font très mal –, ou se contenter de suivre leurs traces sans chercher à les interpréter – ce qu'ils font de mieux en mieux.

Au-dessous du web : la prédiction par les traces

Aussi est-ce vers la seconde solution que s'est orientée la dernière famille de calcul numérique, qui s'est glissée *sous* le web pour enregistrer le plus discrètement possible les traces de ce que font les internautes. Elle se caractérise par l'usage d'une technique statistique particulière, l'apprentissage automatique (*machine learning*), qui est en train de bouleverser la manière dont les calculs pénètrent nos sociétés.

Son ambition est de personnaliser les calculs à partir des traces d'activités des internautes, pour les inciter à agir dans telle direction plutôt que dans telle autre, comme dans les systèmes de recommandation d'Amazon ou de Netflix. Du reste, ces techniques prédictives ont aujourd'hui été ajoutées à la plupart des algorithmes mesurant la popularité, l'autorité ou la réputation. L'algorithme *apprend* en comparant un profil à ceux d'autres internautes qui ont effectué la même action que lui. De façon probabiliste, il soupçonne qu'une personne pourrait faire telle ou telle chose qu'elle n'a pas encore faite, parce que celles qui lui ressemblent l'ont, elles, déjà faite.

Le futur de l'internaute est prédit par le passé de ceux qui lui ressemblent. Il n'est plus nécessaire de trier les informations à partir du contenu des documents, des jugements proférés par les experts, du volume de l'audience, de la reconnaissance de la communauté ou des préférences du réseau social de l'utilisateur. Il s'agit désormais de calculer le profil de l'utilisateur à partir des traces de ses activités, en développant des techniques d'enregistrement qui collent au plus près de ses gestes.

Pour justifier le développement de ces outils prédictifs, les promoteurs des *big data* ont entrepris de disqualifier la sagesse et la pertinence des jugements humains. Les individus, soutiennent-ils, ne cessent de faire des erreurs d'évaluation. Ils manquent de discernement, font des estimations

systématiquement trop optimistes, anticipent mal les effets futurs en préférant toujours le présent, se laissent déborder par leurs émotions, s'influencent mutuellement et ne raisonnent pas de façon probabiliste[13]. À grand renfort de travaux de psychologie et d'économie expérimentales, les architectes des nouveaux algorithmes des *big data* assurent qu'il ne faut faire confiance qu'aux conduites réelles des individus, et non à ce qu'ils prétendent faire lorsqu'ils se racontent sur les très expressives plateformes du web social. Les régularités globales observées sur de grandes masses de traces doivent permettre d'estimer ce que l'utilisateur risque de faire *réellement*. Les algorithmes prédictifs ne donnent pas une réponse à ce que les gens disent vouloir faire, mais à ce qu'ils font sans vouloir vraiment se le dire.

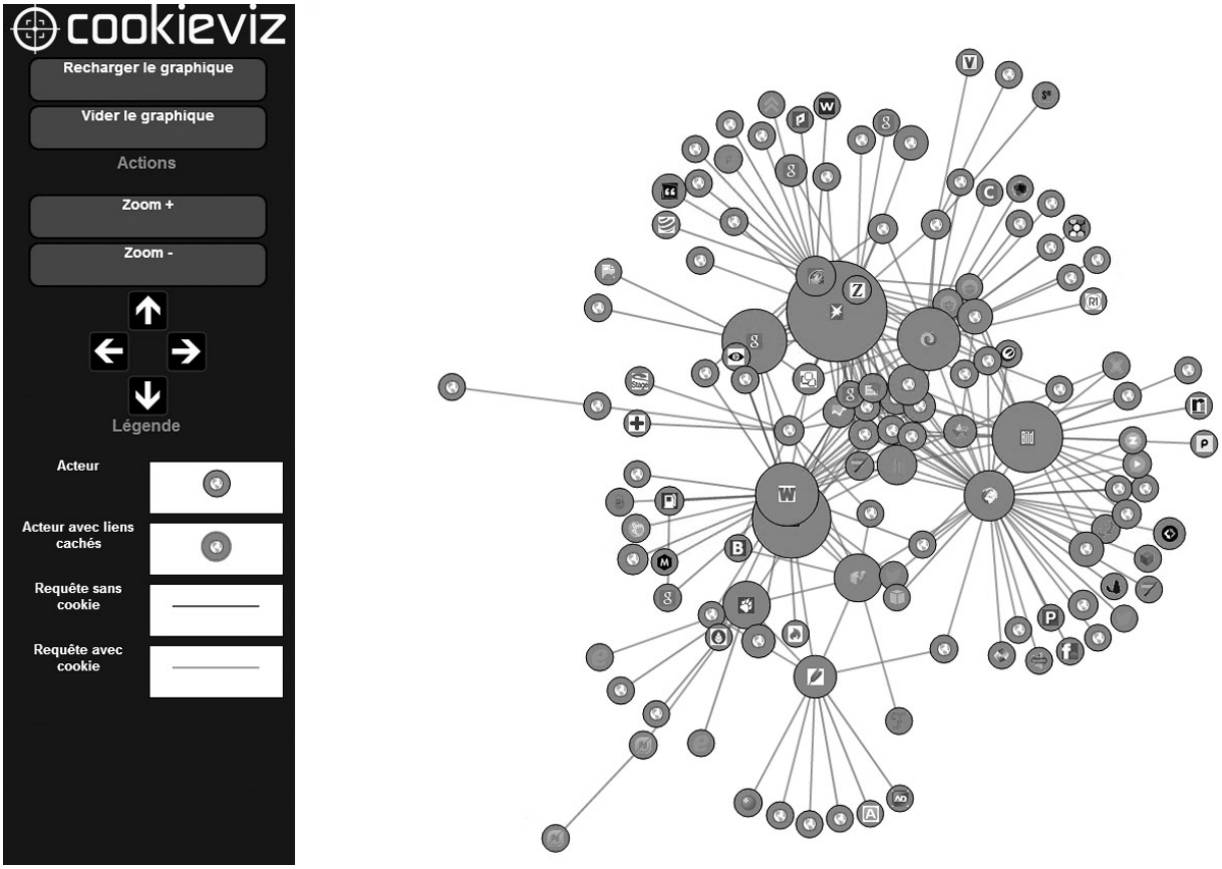
« Je sais que la moitié de mes dépenses publicitaires sont efficaces, mais je ne sais pas lesquelles. » C'est contre cette maxime du marketing traditionnel (mesures d'audience de la première famille) que se déploie aujourd'hui le marketing comportemental de la quatrième famille, en transformant le clic unitaire en une trace longitudinale. Son ambition est de calculer les comportements des internautes pour réduire l'incertitude qui persiste dans les catégorisations trop grossières des segments de styles de vie. Alors que le système d'achat d'espace pour les publicitaires propose d'ajuster les bannières commerciales au contenu éditorial du site, en supposant que celui-ci s'adresse à des catégories sociodémographiques spécifiques, le ciblage comportemental piste l'individu et lui seul.

Le marché de l'affichage publicitaire sur le web est désormais coupé en deux. Les gros sites web parviennent à vendre aux annonceurs des bannières publicitaires sur leurs pages les plus visitées en fonction de l'audience globale. Ils en tirent un prix raisonnable qui s'exprime en CPM (coût pour mille affichages). Mais ce marché de qualité est de plus en plus menacé et tiré vers le bas par le développement de publicités comportementales ciblées qu'affichent automatiquement et à bas prix des réseaux publicitaires sur les pages moins visitées. Ce développement a été rendu possible par la technique des *cookie tiers* : à la différence des *first party cookies*, ils n'appartiennent pas à un éditeur unique qui les dépose dans le navigateur de l'internaute pour le reconnaître lorsqu'il arrive sur son site, mais ils sont la

propriété d'une régie publicitaire en ligne, un *ad-network* comme Weborama, DoubleClick, Critéo ou Right Media.

Lorsqu'un site web confie une partie ou la totalité de la publicité sur son site à un *ad-network*, il autorise la régie publicitaire à profiter des informations de navigation de l'internaute non seulement sur le site qui a servi de cheval de Troie pour introduire le cookie dans le navigateur, mais aussi sur l'ensemble des sites affiliés à cette régie. Mouchard local, le cookie devient alors un espion doté d'un don d'ubiquité. Lorsqu'un internaute s'est inscrit sur Facebook, il suffit qu'il navigue sur un site proposant des « boutons » permettant de partager un article sur Facebook pour que le cookie envoie à l'entreprise de Mark Zuckerberg une information sur son passage sur ce site.

CookieViz[14], une extension du navigateur développé par la CNIL, permet de visualiser les multiples trajectoires que prennent les données de l'utilisateur lorsque, croyant se connecter au site de son quotidien préféré, il est en réalité en train d'envoyer des informations à une dizaine de serveurs obscurs dispersés à travers le monde.



Lecture : En naviguant sur le site représenté par le rond au centre du schéma, l'internaute envoie des informations de navigation aux autres sites et serveurs de données représentés sur ce graphe.
Source : CNIL.

Aux formes traditionnelles de caractérisation de l'internaute dans les catégories du marketing (âge, niveau de vie, revenu) s'ajoutent, jusqu'à s'y substituer, d'autres types d'informations inférées par les calculateurs à partir du pistage de l'internaute. Les régies publicitaires peuvent ainsi prétendre ne pas s'occuper des caractéristiques sociodémographiques des utilisateurs et traiter anonymement les données : de fait, les personnes ont tendance à disparaître derrière leurs traces. La plupart du temps, les éditeurs des sites web ont perdu tout contrôle sur la décision prise par la régie tierce d'afficher telle ou telle publicité sur la page qu'ils proposent à leur lecteur.

Avec la technique du reciblage (*retargeting*) qui domine aujourd'hui ce marché, lorsque le cookie tiers a enregistré qu'un internaute cliquait sur le

site *jetauraiunjour.com*, il va proposer la publicité pour *jetauraiunjour* sur d'autres sites que visitera par la suite l'internaute, et ceci parfois pendant plusieurs semaines. Comme un sparadrap collé au doigt, le *retargeting* est aujourd'hui une des formes dominantes de la publicité numérique. Dans un marché où les revenus publicitaires sont en réalité très bas, il offre à court terme une efficacité légèrement supérieure à une publicité non ciblée, tout en détruisant à long terme la confiance des clients.

Les coulisses du marché de la donnée publicitaire constituent un monde opaque qui entretient une discrétion délibérée, afin de ne pas susciter l'hostilité du public. Les prospères entreprises qui dominent cet univers sont inconnues du grand public (Axciom, BlueKai, eXelate, Rapleaf, Weborama). Ces courtiers de données (*data brokers*) investissent désormais des places de marchés (*ad exchange*) pour s'échanger les données des utilisateurs. Ils compilent les informations aux franges de la légalité, en profitant de la mansuétude des législations. Ils blanchissent leurs activités en se drapant derrière le consentement qu'ils ont extorqué à l'internaute à travers l'acceptation de Conditions générales d'utilisation (CGU) illisibles et jamais lues.

Les publicitaires ont imposé le *cookie tiers* en catimini, en ne suscitant qu'une molle protestation de l'IETF, l'instance de gouvernance des normes techniques d'Internet. Son insertion dans les mondes numériques est attentatoire au principe de respect des données personnelles. Il est spécieux de laisser penser que, lorsqu'un site demande le consentement de l'utilisateur pour son usage, l'utilisateur sait qu'il consent aussi à ce que le cookie espionne ses navigations sur d'autres sites. De surcroît, son emploi a contribué à dégrader le marché publicitaire, en faisant chuter les prix et en conduisant les éditeurs dans une vaine recherche de la parfaite personnalisation[15].

Sans doute est-il aujourd'hui trop tard pour revenir en arrière, mais les critiques qui s'élèvent contre ces mouchards sont de plus en plus vives. En Allemagne, 40 % des internautes utilisent les fonctionnalités offertes par les navigateurs pour bloquer la publicité et se mettre à l'abri des traceurs du web. Ils sont 30 % en France. En réponse, les réseaux publicitaires développent des innovations plus discrètes pour identifier autrement l'utilisateur.

La guerre du traçage vient tout juste de commencer. Avec l'augmentation de l'information du public et la sensibilité de plus en plus grande aux questions de surveillance, consécutive aux révélations d'Edward Snowden, il devient urgent que les régulateurs imposent des règles beaucoup plus dures et exigent des internautes un consentement vraiment éclairé.

Notes

1. Cécile Méadel, *Quantifier le public. Histoire des mesures d'audience de la radio et de la télévision*, Paris, Économica, 2010.

2. Benedict Anderson, *L'Imaginaire national. Réflexions sur l'origine et l'essor du nationalisme*, Paris, La Découverte, 1996.

3. Thomas Beauvisage, « Compter, mesurer et observer les usages du web : outils et méthodes », *in* Christine Barats (dir.), *Manuel d'analyse du web en sciences humaines et sociales*, Paris, Armand Colin, 2013.

4. Alan Ouakrat, Jean-Samuel Beuscart et Kevin Mellet, « Les régies publicitaires de la presse en ligne », *Réseaux*, n° 160-161, 2010.

5. Josiane Jouët, « Les dispositifs de construction de l'internaute par les mesures d'audience », *Le Temps des médias*, vol. 3, n° 2, 2004.

6. Joseph Turow, *The Daily You. How the New Advertising Industry is Defining Your Identity and Your Worth*, New Haven, Yale University Press, 2011.

7. Dominique Cardon, « Dans l'esprit du PageRank. Une enquête sur l'algorithme de Google », *Réseaux*, n° 177, 2012.

8. Matthew Hindman, *The Myth of Digital Democracy*, Princeton, Princeton University Press, 2009.

9. Dominique Pasquier, Valérie Beaudouin et Thomas Legon, *« Moi je lui donne 5/5 ». Paradoxes de la critique amateur en ligne*, Paris, Presses des Mines, 2014.

10. Thomas Beauvisage, Jean-Samuel Beuscart, Kevin Mellet et Marie Trespeuch, « Une démocratisation du marché ? Notes et avis de consommateurs dans le secteur de la restauration », *Réseaux*, n° 183, 2014.

11. Eli Pariser, *The Filter Bubble. What the Internet is Hiding from You*, New York, The Penguin Press, 2011.

12. Hartmut Rosa, *Accélération. Une critique sociale du temps*, Paris, La Découverte, 2010.

13. Ian Ayres, *Super Crunchers. Why Thinking-by-Numbers is the New Way to be Smart*, New York, Random House, 2007.

14. Disponible au téléchargement : http://www.cnil.fr/vos-droits/vos-traces/les-cookies/telechargez-cookieviz/

15. Julia Cagé, *Sauver les médias. Capitalisme, financement participatif et démocratie*, Paris, Seuil/La République des Idées, 2015.

CHAPITRE 2

La révolution dans les calculs

La révolution des *big data* se trouve moins dans l'accumulation des données que dans la manière de les calculer. Chaque famille de calculateurs suppose des conceptions et des usages différents de la statistique. L'arrivée des *data-analystes*, armés de leurs techniques de traitement massif des données, a mis le monde des statisticiens en ébullition. Trois secousses sont venues déplacer la manière dont nos sociétés se représentent à travers leurs chiffres : les mesurés sont eux-mêmes devenus plus facilement calculateurs ; les catégories parviennent moins bien à représenter des individus qui se singularisent de plus en plus ; les corrélations statistiques ne vont plus de la cause vers la conséquence, mais remontent des conséquences vers une estimation des causes probables.

Ces trois secousses ont fait vaciller la longue tradition statisticienne qui s'était constituée, en même temps que les États, pour proposer une cartographie de la nation autour de conventions stables et de catégories de description du monde social[1]. Celles-ci garantissaient, d'une part, un certain degré de consistance et de solidité à travers la notion de moyenne et, d'autre part, une lisibilité suffisante pour offrir des catégories de perception ordinaires.

Or, depuis les années 1980, la société « tient » de moins en moins bien dans les catégories à travers lesquelles les institutions prétendent l'enregistrer, la mesurer et agir sur elle. La crise de la représentation politique est, de façon souterraine, solidaire de l'affaiblissement des formes statistiques qui donnaient une ossature au monde social. La méfiance que les individus manifestent parfois à l'égard des hommes politiques, des

journalistes, des experts ou des syndicalistes a pour soubassement le refus de se laisser enfermer dans des classifications préalablement définies.

C'est précisément pour faire droit à cette revendication de singularité qu'un vaste processus de réinvention des techniques statistiques s'est mis en branle pour calculer la société sans catégoriser les individus. Les nouveaux calculs numériques partent des traces d'activités des personnes, mais ne cherchent pas à en inférer des caractéristiques relatives à des phénomènes plus vastes permettant à la société tout entière de se représenter et de se comprendre. Les catégories statistiques traditionnelles ne déshabillent pas les individus et instaurent des mécanismes de mutualisation des risques pour faire face à l'incertitude des comportements individuels. Désormais, assurent les promoteurs des nouveaux calculs, il va être possible de connaître avec précision les destins individuels et de s'adresser aux individus en s'affranchissant de la solidarité collective.

La manipulation du réel

Une première fragilisation du modèle standard de la statistique sociale s'observe dans le déplacement de la position du calculateur, que l'on a vu migrer quatre fois pour occuper des positions différentes par rapport aux données qu'il mesure.

À côté du web, il s'est occupé à compter les internautes qui cliquent pour agréger des « publics ». Au-dessus du web, il a cherché à se faire oublier des internautes pour dénombrer les marques de reconnaissance qu'ils s'échangent. À l'inverse, c'est en se logeant dans le web lui-même que les métriques du web social sont venues outiller les stratégies de ceux pour qui la visibilité n'est pas un mérite, mais la conséquence d'un travail de production de soi. Enfin, c'est sous le web que, méfiants à l'égard des déclarations intempestives des internautes, les algorithmes sont venus se glisser pour enregistrer les traces de ce qu'ils font réellement.

Ces déplacements montrent comment les statistiques, photographies extérieures de la société, sont progressivement entrées *dans les subjectivités contemporaines* en leur permettant de se comparer, avant de venir subrepticement calculer à leur insu le comportement des personnes. Ce qui

s'observait au-dessus des individus, à travers les catégories qui permettaient de les réunir, s'observe désormais en dessous, par les traces qui les singularisent. Ouvertes, accessibles et destinées à un usage public garanti par l'État, les statistiques sont devenues privées, monopolisées par des entreprises qui les utilisent à des fins commerciales.

Cette trajectoire témoigne du problème de réflexivité que suscite un usage intensif des statistiques par les acteurs du monde social[2]. À la différence du monde naturel observé par la science, la société adapte son comportement aux informations statistiques qui sont données sur elle. L'idéal de l'objectivité instrumentale des sciences naturelles est essentiel pour stabiliser des « faits »[3]. Il dote les objets statistiques de la confiance dont ils ont besoin pour donner des cadres au débat public.

Cependant, dans notre société des calculs, la position d'extériorité de la mesure est de plus en plus difficile à tenir. Les principaux indicateurs de la statistique sociale sont accusés de mal représenter. Sujets de soupçons méthodologiques et de présomptions d'instrumentalisation, ils ont perdu leur aura. Dans les années 1970, la sociologie avait contribué à produire des représentations agrégées des catégories socioprofessionnelles (CSP) permettant de dresser un tableau de la société et de faire apparaître des inégalités dans les trajectoires de mobilité sociale ou dans la distribution de l'accès à la réussite scolaire ou aux biens culturels[4].

Les politiques néolibérales des années 1980 ont contribué à faire perdre leur autorité à ces catégories, en assignant de nouveaux usages aux instruments statistiques : ils servent désormais moins à représenter le réel qu'à agir sur lui. Les techniques de *benchmarking* ont contribué à faire descendre les métriques au sein des mondes sociaux qu'elles prétendent décrire. Le développement du *new public management*, des nouvelles normes comptables dans les entreprises et des dispositifs d'évaluation et de notations a généralisé les index, les palmarès, les baromètres et les *Key performance indicators* (KPI).

Les « vérités » statistiques sont devenues instrumentales : ce n'est plus la valeur propre du chiffre qui importe, mais l'évolution de la valeur mesurée entre deux enregistrements. « Dès qu'une mesure devient un objectif, elle cesse d'être une bonne mesure », souligne la fameuse loi de Goodhart[5].

Mais une autre finalité a été assignée aux indicateurs : rendre les acteurs calculateurs en les insérant dans un environnement qui leur dicte les manières de se mesurer tout en leur laissant une certaine autonomie. Mal reliés entre eux, les indicateurs en batterie ne font plus système. L'expertise calculatoire se substitue à l'autorité professionnelle. Le fait que les mesures soient fausses n'est plus considéré comme un problème[6].

Ce qui importe, en revanche, est d'installer une boucle réflexive qui conduit les acteurs à se savoir sous le regard d'une métrique et à orienter leurs actions en direction des effets qu'elles auront sur la mesure. Les métriques servent à fabriquer le futur. Elles assouplissent nos croyances dans la solidité des chiffres, afin de rendre le réel plus plastique. Celui-ci n'est plus mesuré de l'extérieur mais de l'intérieur, si bien que le périmètre des catégories, au lieu de faire l'objet d'une mesure indépendante, est une variable produite par l'activité de ceux qui sont mesurés : lorsqu'une politique publique enjoint les policiers d'augmenter les chiffres de délits de racolage passif, leurs actions contribuent à faire augmenter le périmètre d'une catégorie qui n'existe pas indépendamment de leurs actions[7].

Il devient ainsi de plus en plus fréquent qu'une mesure d'activité soit prise pour une mesure du phénomène sur lequel s'exerce cette activité : les plaintes des femmes battues deviennent le nombre de femmes battues, les chercheurs les plus cités deviennent les « meilleurs »[8], les lycées qui ont le meilleur résultat au bac sont les meilleures écoles, etc. Un indicateur de performance, souvent unique, devient un outil de lecture d'un contexte bien plus général. La réflexivité des compteurs a non seulement rendu les acteurs de plus en plus stratèges, mais elle a aussi rendu le réel de plus en plus manipulable. Depuis que le site NosDéputés.fr rend public le chiffrage de l'activité des parlementaires, ceux-ci sont non seulement sensibles aux effets de ces classements (notamment parce que la presse locale fait régulièrement le point en titrant « Vos députés sont-ils studieux ? »), mais ils modifient aussi leurs comportements. Les prises de parole des députés en commission se sont multipliées pour augmenter leur score d'activité.

C'est dans ce contexte que les nouveaux calculs des *big data* proposent de réinstaller une position d'extériorité plus solide face aux mesures instrumentales du *benchmarking*. Cependant, ils ne le font plus en cherchant

à se positionner à côté ou au-dessus des données, comme des observateurs distants échantillonnant le monde social depuis leur laboratoire. Ils abandonnent la technique probabiliste du sondage, la vérification de la qualité des données, et cachent le calcul dans la boîte noire des machines, afin que les acteurs ne puissent rétroagir sur elles. Les *big data* réanimient le projet d'objectivité instrumentale des sciences de la nature, mais cette fois sans le laboratoire : c'est le monde qui devient directement mesurable et calculable. Leur ambition est de mesurer au plus près le « réel », de façon exhaustive, discrète et à grain très fin.

Le débordement des catégories

La deuxième secousse qui bouleverse la manière dont la société se reflète dans ses chiffres est la crise des régularités statistiques ordonnant un système stable de catégories entretenant entre elles des liens de dépendance. Les agrégats de la statistique sociale n'accrochent plus sur nos sociétés : ils ne permettent plus ce va-et-vient des individus vers une totalité qui les représente et à laquelle ils s'identifient. Alors que les statistiques n'ont jamais été aussi présentes, elles sont de plus en plus fréquemment contestées. Les indicateurs statistiques globaux, comme le taux de chômage, l'indice des prix ou le PIB, sont souvent identifiés à des constructions statistiques manipulables en fonction des contraintes politiques du moment. Leur rôle dans la figuration du social s'amenuise.

La statistique institutionnelle, qui proposait une articulation entre l'« économique » et le « social » à travers le compromis keynésien et la planification « à la française », a désormais moins vocation à proposer un tableau de la société qu'à mettre en œuvre des politiques gestionnaires. Sous l'influence de la théorie économique des anticipations rationnelles, l'État managérial a réorienté son activité statistique vers des modèles économétriques destinés à anticiper et évaluer les politiques publiques[9]. Au sein des institutions de la statistique nationale, au début des années 1990, la nomenclature des « professions et catégories socioprofessionnelles » (PCS) s'est vue remplacer par des variables plus fines, plus continues et unidimensionnelles comme le diplôme ou le revenu[10].

C'est en effet au nom de la scientificité que l'économétrie a réduit à peau de chagrin les statistiques trop globales des sociologues. À leurs tableaux croisés et à leurs méthodes géométriques d'exploration des données, les économètres préfèrent les techniques de régression linéaire qui procèdent d'une vérification au cas par cas de la corrélation entre deux variables « toutes choses égales par ailleurs ». Les techniques d'analyse des données que les sociologues et les statisticiens avaient établies dans les années 1970, notamment autour des analyses factorielles en composante principale, cherchaient à projeter sur le même plan bidimensionnel un panier d'attributs très variés. Sous l'influence de Pierre Bourdieu, la construction de la variable synthétique de la catégorie socioprofessionnelle permettait de les regrouper pour expliquer ensemble l'origine, la position sociale et les styles de vie des individus[11].

Alors que, via les PCS, les phénomènes sociaux étaient mesurés en les rapportant à un tableau d'ensemble de la société française, le calcul « toutes choses égales par ailleurs » cherche au contraire à isoler, le plus précisément possible, deux phénomènes distincts pour vérifier si l'un agit sur l'autre indépendamment de toute autre variable venue de l'encombrante « société ». Là où l'existence holiste de la société était supposée, elle est mise en suspens pour purifier les interactions entre deux variables. En se mathématisant, le calcul économétrique individualise les données placées en entrée des modèles, en s'efforçant de les rendre les plus fines et univoques possible. Il se méfie des catégories trop englobantes qui risquent de polluer les calculs, de créer des explications tautologiques et de laisser transpirer des présupposés politiques et sociaux.

Le tournant économétrique de la statistique nationale a ainsi préparé le terrain au déploiement des calculateurs numériques des *big data* issus de la physique des grands nombres. Puisque les ressources informatiques le permettent désormais, il n'est plus nécessaire d'affiner les modèles pour assécher la corrélation entre les variables qui lui servent d'hypothèses. Il suffit de demander à la machine de tester toutes les corrélations possibles entre un nombre toujours plus grand de variables. Le modèle n'est plus une entrée dans le calcul, mais une sortie.

Calculer au plus près

Si les calculs deviennent de plus en plus conquérants, c'est aussi parce que la société ne se laisse plus aussi facilement mesurer. La logique de personnalisation renforcée dans laquelle sont entrées les actuelles techniques est une conséquence de l'individualisation expressive qui accompagne le développement des usages du numérique. Dans les sociétés hiérarchiques où l'accès à l'espace public était restreint, il était facile de parler au nom des individus au moyen de catégories qui les représentaient.

Gouvernants, porte-parole ou statisticiens pouvaient « dire » la société à travers les agrégats qu'ils avaient forgés pour la peindre. Aujourd'hui, cette parole abstraite et désincarnée apparaît comme de plus en plus factice et arbitraire. Elle est de moins en moins capable de représenter la diversité des expériences individuelles. Alors que le web a ouvert à tous le droit de prendre la parole en public, le monopole exercé par les représentants sur la description de la société a éclaté et, avec lui, les catégories qui leur servaient à faire parler les autres. La libération des subjectivités ouverte par l'espace public numérique a permis aux individus de s'autoreprésenter[12]. À tous les carrefours de la vie sociale, ils réclament de ne pas être réduits à la catégorie qui les représente. Ils refusent de se laisser enfermer dans les catégories socioprofessionnelles qui ont servi à asseoir une société de statuts. Les patients ne veulent plus être réduits à leur maladie, les clients à leurs achats, les touristes à leurs trajets, les militants à leur organisation, les spectateurs au silence, etc.

L'individualisation des modes de vie et l'augmentation des opportunités sociales ont contribué à augmenter la volatilité des opinions, la diversité des trajectoires personnelles, la pluralisation des centres d'intérêt et la variété des consommations. Un nombre croissant de conduites et d'attitudes se laisse moins immédiatement corréler aux grandes variables explicatives dont sociologues et professionnels du marketing avaient l'habitude. Même si la réalité des déterminations sociologiques des comportements et des opinions est loin d'avoir disparu, elles ne se laissent plus cartographier avec la même évidence que dans *La Distinction* (1979) de Pierre Bourdieu[13], en distribuant un ensemble disparate de goûts et de pratiques politiques,

culturelles, culinaires ou touristiques sur les deux axes de distribution du capital économique et du capital culturel.

La multiplication des pratiques, le développement d'une consommation de loisirs, la diversification des échelles de jugement, les variables d'intensité et de cumul des pratiques ont rendu beaucoup plus complexe – et moins lisible – la distribution des choix culturels[14]. Celle-ci n'a pas disparu, mais, pour l'observer, il faut désormais mobiliser des analyses beaucoup plus sophistiquées et entrer profondément dans les trajectoires de pratiques des individus[15]. Des systèmes de classement qui étaient devenus des catégories ordinaires de perception du monde social ne fonctionnent plus comme des instruments de lecture partagés par tous[16].

Les développements récents des techniques statistiques cherchent à pallier ces difficultés en renouvelant à la fois la nature des données et les méthodes de calcul. Un déplacement systématique s'opère désormais dans le choix des données destinées aux calculateurs. Aux variables stables, pérennes et structurantes, qui fixaient les objets statistiques dans des catégories, les algorithmes numériques préfèrent *capturer des événements* (un clic, un achat, une interaction, etc.) qu'ils enregistrent à la volée pour les comparer à d'autres événements, sans avoir à procéder à une catégorisation. Plutôt que des variables « lourdes », ils cherchent à mesurer des signaux, des conduites, des actions, des performances.

Les nouveaux systèmes publicitaires sur le web sont des automates fonctionnant sur la base d'un système d'enchères en temps réel (*real-time bidding*). Pendant que l'internaute est en train de charger la page web qu'il désire consulter, son profil est mis aux enchères par un automate afin que des robots programmés par les annonceurs se disputent le meilleur prix pour placer leur bandeau publicitaire. L'opération dure moins de 100 millisecondes. Le profil mis aux enchères n'est pas un de ces portraits types du marketing traditionnel. Les informations livrées aux robots des annonceurs sont les traces des navigations antérieures de l'internaute que des cookies ont enregistrées. À la vitesse de l'éclair, les robots des annonceurs vont proposer un prix d'achat en estimant la probabilité que l'internaute clique sur le bandeau publicitaire à partir des données d'activités d'autres internautes. Les promoteurs de ces systèmes automatisés

assurent que la performance de l'affichage publicitaire est de 30 % supérieure lorsqu'il est réalisé par un automate analyseur de traces plutôt que par un *média planner* humain usant de sa connaissance du marché et de sa clientèle.

Dans le monde sans frontière d'Internet, l'affaiblissement du périmètre national rend aussi plus fragile l'échantillonnage des populations par nationalité. Les internautes sont identifiés au moyen d'autres critères mesurables sur le réseau, comme le fait d'être usagers ou consommateurs, ou par l'appartenance à des communautés ethniques, religieuses ou culturelles. En s'affranchissant des cadres nationaux, la plupart des techniques d'échantillonnage, qui permettaient d'étalonner un phénomène au sein d'une population délimitée, se défont. La porte est ainsi ouverte à d'autres formes statistiques que le pourcentage pour représenter les phénomènes mesurés, comme le classement, l'indice et le baromètre, qui compare une activité à elle-même sans la réinscrire dans un tableau plus général et souvent sans connaître le périmètre précis de la population concernée.

L'épuisement des techniques d'échantillonnage a encouragé une rupture radicale dans les méthodologies statistiques. Forts de la puissance de calcul des ordinateurs, les promoteurs des *big data* réclament des corpus complets de données et se réjouissent de les capturer « brutes ». Certains assurent qu'il est préférable de prélever toutes les données sans avoir besoin de les choisir correctement. L'abandon de l'exigence méthodique de sélection des données dans les calculs numériques a plusieurs conséquences.

Les enregistrements ne concernent que les actifs, ceux qui ont laissé des traces ; les autres, non connectés, non agissants, non tracés, sont tout simplement exclus d'une architecture de données en réseau. Ensuite, l'absence d'infrastructure catégorielle permettant de faire tenir les enregistrements statistiques dans un ensemble contribue à la personnalisation des calculs. Dans la plupart des services web mettant en œuvre des techniques de traitements massifs des données, il s'agit de rendre à l'internaute lui-même les informations qui lui correspondent. De façon significative, le seul outil de représentation assurant une généralisation des données est la carte géographique. La géolocalisation permettant de zoomer et de dézoomer sur sa propre situation est le dernier outil de totalisation qui

reste, lorsque toutes les nomenclatures ont disparu[17]. Mon quartier est-il protégé des criminels ? La valeur immobilière de ma rue est-elle en train de croître ? Le service de nettoyage de ma mairie est-il efficace ? Le zoom permet aux individus de se voir dans les données, mais il ne leur permet plus de s'extraire des calculs de leurs traces pour remonter la chaîne d'interdépendance entre le local et le global et découvrir les liens qui font tenir les actions de chacun au système. L'internaute est collé par l'algorithme à ses propres traces sans pouvoir s'en distancier.

Corrélations sans causes

Réplique des deux précédentes, une troisième secousse ébranle les repères de la statistique standard : les corrélations n'ont pas besoin de causes. Dans un article qui a fait grand bruit, Chris Anderson, un des gourous de la Silicon Valley, a annoncé la « fin de la théorie ». Les calculateurs des *big data*, explique-t-il, peuvent désormais chercher des corrélations sans se préoccuper d'avoir un modèle qui leur donne une explication. Les données massives et les mathématiques permettraient de faire l'économie des sciences de l'homme.

« Qui sait pourquoi les gens font ce qu'ils font ? Le fait est qu'ils le font et on peut l'enregistrer avec une fidélité sans précédent. Avec assez de données, les chiffres parlent d'eux-mêmes[18]. » Prenant acte de notre méconnaissance des causes qui sont à l'origine de l'action des individus, les calculateurs abandonnent la recherche d'un modèle permettant de l'expliquer *a priori*. Aussi est-ce un nouveau rapport à la causalité qui s'est mis en place dans certains secteurs de la statistique, conférant aux modèles dits « bayesiens » une victoire posthume sur la statistique « fréquentiste » développée dans la tradition de Quetelet.

L'affaire Target a fait grand bruit aux États-Unis et présente un exemple très simplifié du principe des méthodes d'apprentissage[19]. Cette enseigne de supermarchés dispose, dans sa base clientèle, des informations d'achat de ses clients et connaît grâce à un livre des naissances une partie de ses clientes qui ont déclaré avoir eu un enfant. La méthode des techniques

d'apprentissage consiste à diviser les données en deux corpus différents. À partir des modifications des comportements d'achats effectués par le sous-corpus des femmes dont on sait qu'elles ont eu récemment un enfant, il est possible de trouver des corrélations entre variables d'achats et d'en faire un modèle. Celui-ci est ensuite appliqué à l'autre sous-corpus, afin de *prédire*, parmi les clientes dont on ne sait pas si elles sont enceintes, celles qui le sont peut-être.

L'algorithme a *appris* son modèle à partir du premier sous-corpus pour *prédire* un événement du deuxième sous-corpus. Certaines des corrélations qui permettent de faire cette prédiction sont extrêmement triviales (acheter des produits pour bébé). D'autres, en revanche, peuvent être beaucoup plus inattendues et ne sont souvent que des variables cachées (des *proxies*) d'autres facteurs qui ne sont pas dans les données. Les assureurs, par exemple, auraient constaté dans les données d'achat de leurs clients que ceux qui achetaient des feutres à placer sous les pieds de table et de chaise, pour ne pas rayer leur parquet, avaient un comportement automobile très prudent et qu'ils pouvaient sans risque leur proposer une réduction de prime.

Ce calcul n'est pas « individuel ». Il n'est possible que parce qu'il existe un très important volume de comportements d'achat. La prédiction n'est qu'une estimation statistique et ne présente aucune certitude. Pourtant, c'est cette cliente-là qui, un jour, va recevoir un coupon de réduction pour « femme enceinte » alors qu'elle n'a confié sa grossesse à personne. La « prédiction » semble avoir deviné son intimité. En fait, le calcul a seulement présumé ce qu'elle pouvait être en observant le comportement des autres. À travers un outillage bien plus élaboré, c'est ce principe qui est désormais appliqué à la détection de l'infidélité des clients, des appariements amoureux sur les sites de rencontres, de la récidive judiciaire ou de certaines maladies diagnostiquées préventivement dans les bases de données médicales.

Les modèles statistiques des nouveaux *data scientists* viennent des sciences exactes. De manière inductive, ils partent à la recherche de régularités en faisant le moins d'hypothèses possible. Les capacités de calcul sont désormais si puissantes qu'elles permettent de tester toutes les corrélations possibles sans en épargner aucune au prétexte que l'hypothèse y

conduisant ne serait jamais faite. Il serait cependant trompeur de considérer que ces méthodes sont à l'affût de corrélations « qui marchent » sans se préoccuper de les expliquer. En réalité, elles produisent bien des modèles de comportements, mais ceux-ci n'apparaissent qu'*ex post* et se présentent comme une suite enchevêtrée d'explications dont les variables jouent différemment selon les profils. Il est par exemple possible que, selon les profils, la prédiction d'aimer la chose A dépende dans un cas plutôt de la couleur des yeux, de l'origine sociale et du nombre de déménagements et, dans un autre cas, du fait d'avoir voyagé en Estonie et d'avoir lu les œuvres complètes de Balzac.

À une théorie unifiée des comportements, les calculateurs substituent une mosaïque constamment révisable de micro-théories contingentes articulant des pseudo-explications locales des conduites probables. Ces calculs sont destinés à guider nos conduites vers les objets les plus probables : ils n'ont pas besoin d'être compris et, très souvent, ils ne peuvent l'être. Cette manière inversée de fabriquer le social témoigne du renversement de la causalité opéré par le calcul statistique pour faire face à l'individualisation de nos sociétés et à l'indétermination de plus en plus grande des déterminants de nos actions. Il est en effet frappant de constater que les logiques actuelles des calculateurs cherchent à redonner des cadres à la société, mais, en quelque sorte, à l'envers et par le bas, en partant des comportements individuels pour en inférer ensuite les attributs qui les rendent statistiquement probables.

Notes

1. Emmanuel Didier, *En quoi consiste l'Amérique ? Les statistiques, le New Deal et la démocratie*, Paris, La Découverte, 2009.

2. Wendy Espeland et Michael Sauder, « Rankings and Reactivity : How Public Measures Recreate Social Worlds », *American Journal of Sociology*, vol. 113, n° 1, 2007.

3. Lorraine Daston et Peter Galison, *Objectivité*, Paris, Les Presses du réel, 2012.

4. Luc Boltanski, « Quelle statistique pour quelle critique ? », *in* Isabelle Bruno, Emmanuel Didier et Julien Prévieux (dir.), *Stat-Activisme*, *op. cit*.

5. Marilyn Strathern, « Improving Ratings' : Audit in the British University System », *European Review*, vol. 5, n° 3, juillet 1997.

6. Alain Desrosières, *Prouver et Gouverner. Une analyse politique des statistiques publiques*, Paris, La Découverte, 2014.

7. Isabelle Bruno et Emmanuel Didier, *Benchmarking*, *op. cit*.

8. Yves Gingras, *Les Dérives de l'évaluation de la recherche. Du bon usage de la bibliométrie*, Paris, Raisons d'agir, 2014.

9. Thomas Angeletti, « Faire la réalité ou s'y faire. La modélisation et les déplacements de la politique économique au tournant des années 1970 », *Politix*, vol. 3, n° 95, 2011.

10. Emmanuel Pierru et Alexis Spire, « Le crépuscule des catégories socioprofessionnelles », *Revue française de science politique*, vol. 58, n° 3, 2008.

11. Alain Desrosières et Laurent Thévenot, *Les Catégories socioprofessionnelles*, Paris, La Découverte, 1988.

12. Dominique Cardon, *La Démocratie Internet. Promesses et limites*, Paris, Seuil/La République des Idées, 2010.

13. Paris, Éditions de Minuit, 1979.

14. Olivier Donnat, *Les Français face à la culture. De l'exclusion à l'éclectisme*, Paris, La Découverte, 1994.

15. Bernard Lahire, *La Culture des individus. Dissonances culturelles et distinction de soi*, Paris, La Découverte, 2004.

16. Luc Boltanski et Laurent Thévenot, « Comment s'orienter dans le monde social », *Sociologie*, vol. 6, n° 1, 2015.

17. Dominique Cardon, « Zoomer ou dézoomer ? Les enjeux politiques des données ouvertes », *in* Bernard Stiegler (dir.), *Digital Studies. Organologie des savoirs et technologies de la connaissance*, Paris, FYP Éditions, 2014.

18. Chris Anderson, « The End of Theory : the Data Deluge Makes the Scientific Method Obsolete », *Wired Magazine*, 2008.

19. Eric Siegel, *Predictive Analytics. The Power to Predict who Will Click, Buy, Lie or Die*, Hoboken, John Wiley & Sons, 2013.

CHAPITRE 3

Les signaux et les traces

Les promoteurs des *big data* font preuve d'un optimisme statistique à toute épreuve. Si Internet a libéré les individus du filtre des médias qui les empêchait de s'exprimer, il faudrait désormais libérer les données des fichiers et des modèles qui les cadenassent. Non sans naïveté, ils soutiennent qu'une fois les données brutes « libérées », il suffira de les calculer pour que les vérités mathématiques sous-jacentes au monde social apparaissent et permettent de réduire les erreurs des gouvernants, les approximations de la médecine ou le gaspillage des marchés.

Accessibles, croisées et livrées aux algorithmes, les données pourraient alors, elles aussi, exprimer des choses qui leur étaient interdites ou qui restaient jusqu'alors inconnues en l'absence de mesures objectives. Si notre monde est imparfait, c'est que nous manquons de données pour le corriger.

Les nouveaux gisements de données

Il est vrai que beaucoup d'entreprises et d'institutions disposent de riches bases de données et les exploitent mal. En permettant une plus grande accessibilité à ces données d'exploitation, de services, de réseaux ou de fonctionnements, les politiques d'ouverture des données (*open data*) cherchent à promouvoir les savoirs, les services et la vigilance citoyenne[1]. Les données du trafic téléphonique, des déplacements de bus ou de l'occupation des bornes de Vélib peuvent être introduites dans un nombre important de services tiers, comme le proposent les promoteurs de la « ville intelligente » (*smart city*). Les institutions publiques possèdent des données

qui devraient être accessibles au public pour favoriser le contre-pouvoir vigilant des associations et des citoyens. L'initiative *data.gouv.fr* rend désormais accessible un ensemble de statistiques publiques dans les domaines du logement, de la culture, de l'économie et de l'emploi, qui permettent de nouvelles articulations entre administrations et citoyens.

La recherche scientifique, qui n'a pas attendu le mouvement des *big data* pour concevoir de grandes infrastructures de calcul, se trouve elle aussi stimulée par les nouvelles données numériques. Cependant, il ne sera possible de tirer parti de ces nouveaux gisements qu'en prenant de la distance avec certaines mythologies qui encombrent le discours des promoteurs des *big data*.

Les données brutes n'existent pas. Toute quantification est une construction qui installe un dispositif de commensuration des enregistrements et établit des conventions pour les interpréter. Il faut bien connaître les catégories de la statistique policière pour interpréter les enregistrements des « mains courantes » des commissariats de police et identifier les effets que les changements des consignes ministérielles exercent sur ces enregistrements. Sortis de leur contexte de production et croisés avec d'autres données, ces chiffres risquent de produire plus de contresens que de connaissance.

Par ailleurs, les données ne parlent qu'en fonction des questionnements et des intérêts de ceux qui les interrogent. Très convoitées, les données de l'assurance maladie sont au centre d'enjeux multiples. Actuellement utilisées par les administrateurs de la Sécurité sociale, elles servent à rationaliser les dépenses, par exemple en détectant les médecins qui font de la sur-prescription. Confiées à des associations de malades, elles peuvent aider à faire apparaître des injustices que le milieu médical se refuse à voir. Par exemple, à partir d'une enquête auprès de 9 000 adhérents, Renaloo, une association de malades du rein, a montré que, de façon implicite, la dialyse est plus prescrite aux malades des classes populaires et la greffe aux classes supérieures[2].

Si ces données étaient confiées aux assureurs, comme il en est de plus en plus question, elles permettraient d'ajuster individuellement les primes aux risques, comme les assureurs envisagent déjà de le faire pour l'assurance

automobile, en enregistrant les traces de la conduite, prudente ou risquée, d'une voiture dont les capteurs mouchardent de précieuses informations.

Mais, en dépit de la numérisation croissante des activités, les données ne sont pas facilement accessibles. Si Internet offre un accès à d'importants gisements de données, elles sont souvent peu structurées, prolixes et sans contexte. Les bases de données les plus pertinentes appartiennent aux administrations, aux entreprises et, surtout, aux grandes plateformes du web (Google, Facebook, Amazon). La plupart ont fermé le robinet à données, afin de s'en réserver l'usage ou d'en commercialiser l'accès. Les données de Facebook ne sont pas accessibles, celles de Google sont très partielles, et désormais Twitter fait payer très cher ses archives. Il existe certes d'importantes exceptions, comme Wikipédia ou OpenStreetMap, qui constituent des biens communs accessibles à tous et sont produits par des communautés bénévoles. Mais les intérêts commerciaux des possesseurs de données, les enjeux de protection de la vie privée et les logiques institutionnelles et bureaucratiques ne cessent de freiner le processus d'ouverture des données.

Enfin, les bases de données numériques sont souvent mal catégorisées et pleines de « bruit ». La plupart du temps, le croisement de données issues de bases hétérogènes est impossible ou demande un délicat travail d'interopérabilité qui menace, parfois, la vie privée des personnes. Persuadés que la quantité peut se substituer à la qualité, les zélotes des *big data* assurent qu'un monde plus mesurable deviendrait aussi plus calculable. Si elles peuvent jouer un rôle considérable dans la transformation de certains secteurs d'activités, il arrive aussi que les mégadonnées produisent plus de bruit que de signal, qu'elles soient biaisées, se trompent ou produisent des résultats indésirables. Aussi faut-il entrer dans le fonctionnement des calculateurs pour comprendre ce dont ils sont – ou ne sont pas – capables.

Des machines « statistiques »

Nous sommes habités par l'idée anthropomorphe que les machines calculatoires seraient intelligentes et que leurs concepteurs seraient parvenus

à glisser un esprit à l'intérieur de leurs mécanismes. Cette conception nourrit nos représentations et nos craintes. De HAL, l'ordinateur fou de *2001, l'Odyssée de l'espace*, aux « précogs mutants » de *Minority Report*, qui prédisent un crime avant qu'il ait eu lieu, elle est constamment alimentée par l'imaginaire de la science-fiction. Pourtant, dans les laboratoires de recherche, personne ne croit vraiment que les algorithmes aient ce type d'intelligence.

Dans les années 1980, le programme de l'Intelligence artificielle (IA) visait à faire reproduire aux automates le raisonnement humain en les dotant de règles, de modèles cognitifs, d'ontologies et de syntaxes reproduisant la complexité des formes logiques et symboliques de la pensée. Ce programme a échoué dans les années 1990 pour de nombreuses raisons, dont la principale est l'incapacité des automates « intelligents » à interpréter l'infinie variété des situations et des contextes[3]. Qu'un ordinateur, *Deep Blue*, ait pu en 1997 battre Garry Kasparov aux échecs constitua l'apothéose et la fin des ambitions de ce programme de recherche. L'intelligence artificielle s'est fracassée contre l'infinie diversité des contextes.

Rendre la machine « intelligente » ne sert à rien si elle ne sait pas adapter son raisonnement à chaque situation. Or la plupart des situations de la vie réelle ne sont pas « codées », comme le sont les règles de déplacement des pièces aux échecs. Aujourd'hui, le projet de l'intelligence artificielle a refait surface dans les projets de recherche en informatique. Il est au cœur des techno-utopies transhumanistes des gourous de la Silicon Valley, comme Ray Kurzweil, à qui Google a confié la responsabilité de son équipe de recherche la plus prospective. Cette nouvelle intelligence artificielle n'a plus grand-chose à voir avec le projet initial. Désormais, les machines cherchent beaucoup moins à modéliser le raisonnement qu'à ingurgiter des contextes à travers d'énormes masses de données. Les concepteurs ont abandonné l'ambition de faire des machines « intelligentes ». Ils préfèrent les rendre « statistiques ».

Pour illustrer ce changement de paradigme, prenons le cas de la traduction automatique. Dans les années 1980, ingénieurs et linguistes ont cherché à mettre dans les programmes des règles de grammaire et de syntaxe abstraites, des dictionnaires de mots et de concepts (appelés

« ontologies »), afin que les traducteurs automatiques puissent reproduire le raisonnement formel permettant le passage d'une langue dans une autre. En dépit d'efforts de recherche considérables, cette stratégie n'a fait faire à la traduction automatique que des progrès fort limités.

IBM puis Google ont alors orienté le projet dans une direction différente. Au lieu de concevoir une machine au raisonnement abstrait, ils ont décidé de lui faire apprendre mot par mot, groupe de deux mots par groupe de deux mots, puis de trois mots, etc., les correspondances entre deux textes déjà traduits par d'autres, comme l'immense corpus réalisé par les traducteurs humains des institutions européennes. La machine ne traduit pas : elle calcule l'estimation statistique de la meilleure traduction de ces deux (trois, quatre, etc.) mots, en les comparant à toutes les autres traductions qu'elle a en mémoire.

Pour « apprendre », l'ordinateur a donc besoin d'avaler le plus gros tas de textes possible et de leurs traductions dans les langues visées. La machine ne « comprend » rien de ce qu'elle fait, mais, en s'appuyant sur la masse considérable de données qui lui fournissent des milliers de contextes d'utilisation de différents sacs de mots, elle peut estimer les correspondances statistiques les plus probables dans une autre langue. La qualité de la traduction de Google Traduction est loin d'être optimale, mais ce changement de paradigme lui a permis de faire des progrès considérables. Dans beaucoup d'autres domaines d'aide à la décision (médicale, juridique, financière, diagnostic technique, etc.) où il existe un savoir codifié, les « systèmes experts » ont ainsi beaucoup gagné de l'apprentissage statistique. Au moyen d'une tout autre technique d'apprentissage (appelée *deep learning*), la reconnaissance vocale ou la détection de formes dans les images font elles aussi des progrès considérables aujourd'hui.

Pour une large part, l'innovation des *big data* réside dans ce passage des règles abstraites vers la statistique des contextes. L'enjeu n'est plus d'apprendre aux machines une grande théorie appliquée à peu de données, mais de multiplier les petites théories en demandant à beaucoup de données contextuelles de sélectionner la ou les meilleures d'entre elles. Les capacités de calcul désormais disponibles permettent de tester plusieurs milliers d'hypothèses en même temps.

Mais ceci n'empêche pas de prendre en compte plusieurs théories dans la prédiction ni, surtout, de changer les pondérations affectées aux différentes hypothèses pour chaque profil et chaque contexte d'utilisation. Les méthodes non paramétriques ont ceci d'agnostique qu'elles ne figent pas la contribution de leurs variables, mais les révisent constamment en fonction des actions de l'utilisateur. Pour cette raison, il est vain de réclamer que soit levé le « secret » des algorithmes et plus utile de connaître les flux de données qui « entrent » dans la composition du calcul. Ceux qui les fabriquent ne savent pas eux-mêmes expliquer pourquoi le calculateur a, dans ce contexte, retenu telle hypothèse plutôt que telle autre.

En revanche, la plupart des méthodes d'apprentissage sont dites « supervisées » : ceux qui fabriquent les calculs leur donnent un objectif. Pour choisir les meilleurs contextes parmi les données disponibles, il sera demandé à l'algorithme de maximiser l'efficacité du service rendu à l'utilisateur : qu'il passe le plus de temps sur Facebook, qu'il clique le plus fréquemment sur les recommandations, qu'il exploite le premier lien des classements du moteur de recherche. Qu'on lui confie des données et un objectif, l'algorithme définira lui-même la bonne théorie pour que, en chaque situation, les corrélations soient les plus efficaces par rapport à l'objectif visé.

S'il n'est guère possible d'enquêter dans les variables versatiles des algorithmes, il est en revanche décisif de demander à ceux qui les fabriquent de rendre publics les objectifs qu'ils leur donnent. Une des particularités de ces modèles « auto-apprenant » est que les internautes sont constamment soumis à des expérimentations sans le savoir. De façon massive, les grands services du web proposent des versions différentes de leurs services à des groupes d'utilisateurs différents, afin de tester et de comparer les hypothèses qui nourrissent leurs objectifs. Le *A/B testing*, cette technique d'échantillonnage en double aveugle, habituellement utilisée à titre expérimental, conçoit désormais la société comme un laboratoire à grande échelle. Nous sommes leurs cobayes.

Le signal et la trace

Dans les services numériques, un algorithme « fonctionne » véritablement lorsqu'il parvient à épouser si étroitement le milieu dans lequel il intervient que les comportements des acteurs se règlent sur ses verdicts et que les principes qu'il met en œuvre nourrissent leurs représentations. On peut le dire du PageRank de Google, du système de recommandation d'Amazon, des notes d'hôtel de TripAdvisor ou du GPS embarqué dans les voitures.

En revanche, beaucoup d'autres services calculés ne parviennent pas à produire des résultats suffisamment intelligibles pour redéfinir les mondes sociaux dans lesquels ils interviennent. Les services d'analyse de sentiment, qui promettent de dégager la tonalité subjective de grandes quantités de textes numériques, produisent des résultats si triviaux que leurs utilisations se limitent à produire une représentation très approximative des opinions positives ou négatives concernant une marque, un produit ou une personnalité.

Pour expliquer ces différences, il est nécessaire de séparer, au sein des proliférantes *big data*, les données qui proposent des contenus explicites, informations ou expressions subjectives – appelons ces données des *signaux* (par exemple un statut sur Facebook) – et celles, implicites, qui sont des enregistrements contextuels de comportements – appelons ces données des *traces* (clics, géolocalisation, navigation, vitesse de lecture, etc.). Les algorithmes du web les plus « efficaces » sont ceux qui couplent étroitement des *signaux informationnels* avec des *traces de comportement* ou, pour le dire autrement, qui se servent des *traces* pour trouver la meilleure relation entre les *signaux*. En revanche, lorsque les calculs sont appliqués sur des signaux sans traces ou sur des traces qui se réfèrent peu à des signaux, ils n'ont pas la même efficacité.

Les services qui parviennent à créer une boucle d'apprentissage entre *signaux* et *traces* ont pour caractéristique de traiter en temps réel une information qui ne se trouve pas dans les règles de calcul de l'automate, mais qui est logée, sous le web, dans la conduite de l'utilisateur. Avec les techniques d'apprentissage, les algorithmes délèguent la sélection de l'information la plus pertinente (pour leur calcul) aux agissements réguliers des internautes. Cette information peut être extrêmement triviale. Taper « mal à la tête », « fièvre » ou « courbature » dans le moteur de recherche

désigne de façon plausible un symptôme grippal que le géant de Mountain View sera ensuite capable de mesurer et de cartographier, pour proposer une prédiction de l'épidémie de grippe plus rapide que les institutions sanitaires. En réalité, les analyses de *Google flu* montrent que les prédictions de Google sont imprécises et parfois fausses[4].

Longtemps, Google a fait la guerre à ceux qui cherchaient à obtenir de la visibilité en produisant des liens hypertextes sans autorité, c'est-à-dire en envoyant au calculateur du « bruit » sans signal. Désormais, les concepteurs de l'algorithme font de plus en plus confiance à un autre moyen pour parfaire son calcul. Il enregistre le lien sur lequel a cliqué l'internaute parmi la liste de réponses proposées à un mot-clé. Si l'utilisateur ne revient pas sur la page pour cliquer sur un autre lien, l'algorithme va conclure que l'internaute est satisfait par la réponse apportée. Cette information sur le comportement de l'utilisateur permet à l'automate de réviser son classement. S'il apparaît que beaucoup de personnes cliquent systématiquement sur le troisième lien plutôt que sur les deux premiers, il va essayer de comprendre quel est le profil de ce groupe de personnes (ont-ils l'habitude de cliquer sur des liens commerciaux, universitaires, divertissants, francophones ?) et se servir de cette information pour optimiser les résultats.

Les outils de recommandation utilisés pour les livres (Amazon), les films (Netflix), les musiques (Deezer, Spotify) ou des produits qui ressemblent à ceux qui viennent d'être achetés reposent sur une technique appelée « filtrage collaboratif ». Ils proposent à l'utilisateur d'élargir l'offre proposée par des recommandations, en comparant son profil comportemental avec d'autres utilisateurs ayant aimé (ou acheté) les mêmes produits. L'algorithme fait confiance à la régularité des structures de goûts et d'intérêts des utilisateurs pour rendre prévisibles les rapprochements entre les produits recommandés.

Il est possible d'ajouter aux calculs beaucoup d'autres informations. Les algorithmes musicaux, par exemple, emploient aussi des catégories sémantiques pour réunir ensemble les artistes d'un même genre ou sous-genre. Certains essaient de trouver des similitudes entre les morceaux à partir de la forme de leur signal sonore. Mais aucune de ces solutions n'a plus de poids que cette règle simple et efficace : quelqu'un qui écoute A et B

a des chances d'aimer C et D, s'il existe beaucoup d'autres individus qui écoutent A, B, C et D. Les outils de recommandation sont performants parce qu'ils font l'hypothèse qu'il existe un *caractère régulier et prévisible* des pratiques de lecture, d'achat ou d'écoute.

Mieux encore, ces calculs qui se greffent directement sur le comportement de l'utilisateur savent si un morceau de musique a été écouté plusieurs fois (ou passé rapidement), mesurent le temps de lecture (ou d'abandon) des livres sur les tablettes numériques, affinant ainsi leurs modèles d'un savoir précis des gestes de l'utilisateur. Les déplacements tracés par le GPS, les achats par la carte bancaire, les consommations culturelles, les expressions politiques ou les clics de lecture, ces signaux réguliers qui s'appuient sur de solides régularités sociologiques permettent aux systèmes auto-apprenant de réviser leurs règles.

Même en ayant le sentiment de faire des choix singuliers, nos comportements obéissent à des habitudes routinières profondément inscrites dans notre socialisation. C'est pour cela que les infrastructures de calcul qui couplent étroitement des signaux avec des traces monotones produisent des recommandations pertinentes lorsqu'elles disposent d'un volume suffisant de données. Les calculs font travailler les individus pour trouver la bonne théorie. Si Chris Anderson peut soutenir que les nouveaux algorithmes du web laissent se disputer des milliers de corrélations sans causes, c'est parce que leurs concepteurs font implicitement une hypothèse décisive : ils demandent au caractère régulier et monotone du comportement des utilisateurs de stabiliser les modèles en les aidant à apprendre les bonnes corrélations.

Il y a quelques raisons de trouver que les algorithmes fonctionnent bien lorsqu'ils sont sociologues. Pour Pierre Bourdieu, l'*habitus* est cette disposition incorporée à travers laquelle la société façonne des choix réguliers et prévisibles, jusque dans les plus petites anfractuosités du quotidien. Il est à peine inconvenant de dire que les automates fonctionnent en faisant confiance à l'algorithme de leurs utilisateurs, leur *habitus*. La plupart du temps, les prédictions algorithmiques ne font que confirmer, en leur donnant une amplitude plus ou moins grande, des lois sociales bien connues.

PredPol est un service prédictif qui désigne aux polices anglaises et américaines de certaines villes les zones à patrouiller. La prédiction de délits s'appuie sur les résultats des enquêtes de victimation : ce sont souvent les mêmes personnes, dans les mêmes lieux, qui subissent les délits et les crimes venant des mêmes délinquants. Ayant appris l'histoire criminelle d'une ville et en y ajoutant toute une série de variables sur le tissu urbain (heures d'ouverture des magasins, propriétés immobilières, ensoleillement, etc.), l'algorithme produit des prédictions convaincantes. Les policiers à l'ancienne, connaisseurs de leur terrain et habitués aux maraudes, ne sont jamais vraiment surpris des directives du logiciel[5].

Un comportementalisme radical

Est-ce que les algorithmes « marchent » parce que les individus sont réguliers ou les prescriptions des algorithmes les rendent-ils réguliers ? Une récente enquête sur le fil d'actualité (*newsfeed*) de Facebook permet d'éclairer cette question. Sur Facebook, l'utilisateur ne voit pas défiler dans son fil d'actualité toutes les informations publiées par ses amis. Celles-ci sont filtrées par un algorithme, le *EdgeRank*. Son principe est de privilégier les informations publiées par des amis avec lesquels l'utilisateur a fréquemment interagi. Lorsqu'un internaute commente, *like* ou consulte régulièrement la page d'un ami, il verra les publications de cet « ami » apparaître dans ses actualités. En revanche, les publications de ceux avec lesquels il interagit peu ou pas lui seront moins fréquemment montrées, jusqu'à disparaître totalement de l'attention de l'internaute.

L'algorithme propose de s'appuyer sur les pratiques de sociabilité des utilisateurs, en privilégiant les informations de ceux qui ont entre eux une conversation régulière et en laissant dans l'ombre ceux avec lesquels les relations sont distantes et fragiles. Les marques commerciales qui voudraient glisser leurs messages dans la conversation des internautes se plaignent de cette disparition. La plupart du temps, leur page a été « *likée* » à l'occasion d'un quiz ou d'un jeu concours, puis complètement oubliée par les internautes. Aussi le modèle publicitaire de Facebook consiste-t-il à faire payer les marques pour que des « posts sponsorisés », qui, selon les

principes de l'algorithme, ne devraient pas apparaître dans le flux de l'internaute, soient cependant poussés jusqu'à lui.

Facebook gagne ainsi de l'argent en permettant à certains de forcer la porte d'une conversation « privée » que son algorithme essaie de protéger. Beaucoup d'autres facteurs entrent désormais dans la confection du *newsfeed*, comme les liens que l'utilisateur a *liké* sur Internet, certains liens partagés par des amis qui connaissent une forte popularité, le temps de défilement de la page, etc. Avec le développement des techniques de *machine learning*, Facebook substitue de plus en plus au modèle initial de son algorithme (privilégier les amis avec lesquels les interactions sont les plus nombreuses) un apprentissage : reconnaître, parmi toutes les informations publiées, celles que clique le plus l'internaute – en donnant comme objectif à l'automate d'augmenter le temps passé sur la plateforme.

Comme pour beaucoup de mesures de réputation, il est fréquemment fait reproche à Facebook de placer l'utilisateur dans une « bulle » (*filter bubble*) : selon les affinités de l'utilisateur, l'algorithme ferme la fenêtre sur le monde en réduisant son paysage aux choix de ses amis. À partir d'un échantillon de 10 millions de comptes américains, les chercheurs de Facebook ont essayé de mesurer les effets de la bulle en s'intéressant à l'une de ses dimensions : le fait de ne pas « voir » les informations venant d'un bord politique qui n'est pas le sien[6]. Même si l'échantillonnage de l'enquête mérite discussion, les résultats sont très robustes.

Ils montrent que les utilisateurs ferment eux-mêmes leur fenêtre de visibilité plus que ne le fait Facebook en filtrant l'information à travers son algorithme. Si aucun filtre ne s'exerçait, les utilisateurs libéraux verraient dans leur fil 45 % d'informations conservatrices parmi les liens partagés et les conservateurs 40 % d'informations libérales. Mais, en choisissant leurs amis, les utilisateurs donnent une couleur politique particulière à leur sociabilité. Les libéraux n'ont que 18 % d'amis conservateurs et les conservateurs n'ont que 20 % d'amis libéraux. C'est pourquoi les libéraux ne voient en réalité que 24 % d'informations conservatrices et les conservateurs 35 % d'informations libérales. Le filtre de l'algorithme réduit cette proportion, mais ce sont les utilisateurs qui, à travers les liens sur lesquels ils choisissent de cliquer, referment la bulle sur eux-mêmes. Les

libéraux lisent des informations libérales et les conservateurs des informations conservatrices.

Un tel résultat ne surprendra pas les sociologues qui travaillent sur la consommation d'informations politiques. Il conforte le fait que la sociabilité des individus, surtout des plus politisés, est homophile : ils ont, très majoritairement, des amis qui ont les mêmes opinions et valeurs ; ils s'exposent prioritairement à des sources d'information qui confortent leurs idées. Il n'est pas besoin que l'algorithme enferme les individus dans une bulle, ils le font d'eux-mêmes en obéissant aux régularités comportementales qui sont inscrites dans leur socialisation. Si les résultats de cette enquête semblent dédouaner Facebook de biaiser, filtrer ou censurer l'information, ils invitent à déplacer et à formuler autrement la critique qui doit être adressée au guidage algorithmique.

C'est en effet le comportementalisme radical des nouvelles techniques de calcul qu'il faut questionner. Avec une insistance provocante, les concepteurs des algorithmes prédictifs ne cessent de dire qu'ils ne font que s'appuyer sur les comportements passés de l'internaute pour lui recommander des choses à faire. L'utilisateur est ainsi constamment renvoyé à sa seule responsabilité et à ses ressources sociales et culturelles. S'il a des goûts culturels variés, par exemple s'il écoute à la fois John Cage, Beyoncé et de la musique éthiopienne, l'algorithme de Deezer l'aidera à explorer ces différents univers musicaux, même si certains d'entre eux correspondent à des comportements de niche. Si, en revanche, il n'écoute que Beyoncé, l'algorithme lui recommandera les titres les plus *mainstream* de la musique R'n'B.

Les nouvelles techniques de calcul ont ceci de particulier que, à la différence des mesures d'audience ou d'autorité, elles ne cherchent pas à ramener l'individu vers le centre de la société et sa moyenne normée. Si les traces que l'internaute livre à l'algorithme présentent un profil singulier, original ou périphérique, les recommandations qui lui seront faites ne seront pas tirées vers le milieu par des personnes ayant des goûts conformes et communs. Le paradoxe des nouveaux calculs est que, refusant la prescription paternaliste des médias, les individus désormais calculés à travers leurs traces ont des conduites régulières.

Le comportementalisme radical joue ce rôle à la fois lucide et démoralisant de montrer à des sujets qui pensaient s'être émancipés des déterminations que, en ce qu'ils pensent être des singularités inassignables, ils continuent à être prévisibles, petites souris mécaniques dans les griffes des calculateurs. Vue depuis les algorithmes, la société ne repose plus sur de grands systèmes de déterminations, mais elle est une sorte de micro-physique des comportements et des interactions que des capteurs placés à bas niveau savent décoder[7]. Nourris par les sciences de la nature, ceux qui promeuvent ces outils sont persuadés qu'il existe dans le social quelque chose de déterminé et de calculable, si on veut bien l'attraper par le bas, à la manière d'interactions entre atomes, et non par le haut, comme des groupes sociaux en rapport les uns avec les autres.

Il ne suffit pas que l'utilisateur ait envie d'écouter autre chose ou de découvrir des choses qu'il ne connaît pas : il faut que, par ses actes, il prouve concrètement à l'algorithme qu'il souhaite échapper à ses régularités. Une étude sur les utilisateurs d'un site de location de vidéo en ligne australien compare les films que les utilisateurs mettent dans leur liste de souhaits et les films que les utilisateurs regardent réellement sur la plateforme[8]. Elle met en évidence le fait que les individus voudraient regarder des documentaires d'information et des films d'auteur, mais qu'en réalité ils consomment des films grand public à succès. La logique algorithmique colle à ce que font les individus en considérant, de façon très conservatrice, qu'ils sont rarement à la hauteur de leurs désirs. En préférant les conduites aux aspirations, les algorithmes nous imposent ce réalisme efficace. Ils nous emprisonnent dans notre conformisme.

Faut-il proposer des publicités pour les cigarettes à un fumeur qui voudrait arrêter de fumer, sous prétexte qu'il ne parvient pas à mettre en pratique ses résolutions ? Les algorithmes qui se disent prédictifs ne le sont pas parce qu'ils seraient parvenus à entrer dans la subjectivité des personnes pour sonder leurs désirs ou leurs aspirations. Ils sont prédictifs parce qu'ils font constamment l'hypothèse que notre futur sera une reproduction de notre passé. L'individu des algorithmes est un « dividu », selon l'expression que Gilles Deleuze avait forgée pour imaginer la disparition de l'individu pris dans les flux du contrôle machinique[9]. Il n'a pas d'histoire, pas

d'intériorité, pas de représentations ni de projets. Il n'est pas inscrit dans une position, pris dans des rapports sociaux, soumis aux forces multiples qui s'exercent sur lui. Il est ce que trahissent ses comportements dans le miroir que lui tendent les autres. Le comportementalisme algorithmique, c'est ce qui reste de l'*habitus* lorsqu'on a fait disparaître les structures sociales.

Le couplage algorithmique des signaux et des traces n'est cependant efficace que dans des conditions bien particulières. La boucle de rétroaction entre l'information et l'action qui la valide est toujours extrêmement courte : proposer des livres, les acheter ; classer des liens, les cliquer ; indiquer un parcours, le suivre. Lorsque le couplage est plus long, plus incertain, lorsqu'il donne plus de place à une réflexion distante de l'utilisateur, les calculs prédictifs sont à la peine.

Par ailleurs, ils doivent disposer d'un volume considérable de traces de comportements afin de brasser suffisamment de contextes pour calculer les choix les plus singuliers. C'est pourquoi les grands systèmes de calculs actuels capturent les flux de données et les utilisateurs au sein de grandes plateformes mondiales (Apple, Facebook, Amazon, etc.). Les données sont enfermées dans les boîtes noires de machines d'une rare complexité. En temps réel, les signaux et les traces sont sans cesse calibrés, testés et alignés à l'intérieur d'écosystèmes fermés, et non pas croisés au grand air avec des flux d'informations venant de toutes parts. Il est à craindre que le monde des *big data* encourage beaucoup plus la fermeture sur de grands monopoles industriels que le tissu diversifié d'acteurs se partageant de grands stocks de données ouvertes, ce dont rêvent les promoteurs de l'*open data*.

Signaux sans traces et traces sans signaux

Beaucoup de projets des *big data* ne cherchent pas à coupler signaux et traces dans une boucle qui se referme sur les comportements de l'utilisateur. Certains rassemblent de nombreux signaux mais peu de traces, alors que d'autres enregistrent des traces mais n'ont pas beaucoup de signaux. Pour rendre intelligibles ces données, les calculateurs doivent emprunter d'autres voies et ce sont alors d'autres représentations de la société qu'ils façonnent.

L'explosive expressivité numérique a mis en circulation sur le web un nombre considérable de tweets, de posts, de blogs, de photographies, de *selfies*, de *check-in* et de collections d'informations de toutes origines et de toutes sortes. Cette prolifération de données non structurées et sans contexte, aléatoires et contingentes, bavardes et explosives, redondantes et prolixes, rend disponible une masse considérable de signaux originaux pour représenter les pulsations de la vie numérique. Cartographie mondiale de la diffusion du hashtag #JeSuisCharlie, palmarès des mots-clés les plus recherchés dans Google, réseaux de la blogosphère politique, circulation virale d'une vidéo à succès, représentation d'une polémique sur Twitter… De nombreuses techniques de représentation s'attachent à visualiser la profusion de ces signaux du web : graphes interactifs, nuages de mots-clés, cartographies lumineuses, frises chronologiques, courbes, histogrammes et camemberts, murs de photos zoomables, listes et classements en tous genres, etc.

Les calculs opérés à partir des mesures de réputation numérique produisent des visualisations séduisantes et attractives. Ils fournissent de nouveaux formats d'écriture au data-journalisme, permettent aux internautes de se saisir des données des autres pour les valoriser et ouvrent de nouvelles voies aux méthodes digitales des sciences sociales[10]. La représentation de ces données a cependant perdu toute référence à l'idée de représentativité statistique. Sans population à laquelle les comparer, les signaux numériques parlent avant tout pour eux-mêmes.

En premier lieu, les signaux qui sont envoyés sur le web par les internautes abritent des stratégies de mises en visibilité qui valorisent les plus actifs. Comme toutes les métriques de réputation situées dans le web, il est facile de manipuler ces indicateurs. Les émissions de télévision sont devenues attentives aux conversations qu'elles suscitent et encouragent sur Twitter. Les marques mesurent leur réputation en dénombrant le nombre de *like* qu'elles ont reçus (ou achetés) sur Facebook. Les militants politiques s'installent dans des *war rooms* pour bombarder Twitter de messages quand leur candidat passe à la télévision. Les commerçants suscitent de faux commentaires pour répondre aux vrais qui pourraient ruiner leur réputation.

Deuxièmement, ces signaux très inégalement produits par les internautes n'ont comme seule trace de contexte qu'un horodatage et une

géolocalisation. La plupart des représentations dont ils font l'objet les superposent sur une carte ou une ligne temporelle. Les données expressives du web mettent en circulation de la connaissance, de l'influence, de la critique ou de la réputation. Les calculateurs du web les visualisent avec brio et originalité, mais peinent à les interpréter. Le statut qui doit être donné à ces signaux constitue aujourd'hui un enjeu à la fois épistémologique et politique.

Ces mots-dièses partout répétés valent-ils une opinion ou sont-ils juste l'effet mimétique d'une contagion sans importance ? Ces protestations qui se coagulent autour d'une page Facebook méritent-elles d'être prises pour un engagement politique ? Le nombre de fans qu'une entreprise capitalise sur sa page Facebook a-t-il une quelconque valeur ? Les signaux numériques qui, partout sur le web, expriment, agitent et témoignent donnent à voir la société en mouvement, mais ils ne peuvent être assignés à la hiérarchisation des enjeux publics à laquelle les médias et les sondages nous ont habitués avec leurs mesures de représentativité. La société numérique ne parle plus à son centre.

Privés de contexte, les calculateurs doivent trouver dans les signaux numériques des prises pour les représenter. Les technologies de traitement automatique de la langue (TAL), d'analyse des sentiments et les démarches du web sémantique cherchent à extraire des significations de ces textes, images et vidéos pour les comprendre et les interpréter. Les projets de recherche dans le domaine sont considérables et beaucoup de start-up se sont lancées dans l'idée de comprendre et de résumer le contenu des expressions du web. Les résultats sont à la fois incertains et triviaux[11]. Les analyses textuelles parviennent à ranger l'information, à catégoriser les contenus autour de thèmes, à montrer la circulation d'arguments, à produire des catalogues, mais elles sont incapables de saisir correctement les énonciations.

Beaucoup promettent, sans résultats, d'extraire de cette profusion de signaux la prédiction d'événements futurs : les résultats des élections, la réussite d'un film, la carrière d'un chanteur ou le cours de Bourse d'une entreprise. Souvent citée en exemple, la société Epagogix serait parvenue à créer un outil de prédiction du succès des films. Ayant en mémoire l'ensemble des scénarios hollywoodiens, le nom et le cachet des acteurs et

metteurs en scène, connaissant le nombre d'entrées en salles et l'audience des diffusions télévisées de chaque film, on pourrait tout prévoir : il suffirait qu'un jeune cinéaste débutant présente son scénario à la machine pour qu'elle lui prédise les recettes du film, indique les acteurs à recruter et, éventuellement, les points du scénario à modifier[12].

C'est avec ce genre de promesses improbables que la fortune des *big data* fait son chemin chez les consultants, les stratèges d'entreprises et dans les médias. En dépit de l'amplification des mesures, nous ne savons pas mieux prévoir les crises financières, les tremblements de terre, les scores des matchs de football et les résultats électoraux[13]. Lors du référendum sur l'indépendance de l'Écosse, l'analyse des tweets donnait presque 60 % de votes pour le oui. Il est possible de faire de nombreuses prédictions à partir des données du web, mais, comme toute prédiction, elles sont des estimations statistiques imparfaites. Dans le monde de la publicité personnalisée, on aime à dire que la « performance d'une campagne a été augmentée de 100 % grâce à un algorithme ». Mais cette victoire signifie généralement que le nombre d'utilisateurs qui cliquent sur le message publicitaire est passé de 0,01 % à 0,02 % ! En l'absence de culture statistique, nous avons tendance à penser les estimations statistiques comme des assertions définitives, certaines et paralysantes. En réalité, elles ne font que dire, avec plus ou moins d'approximations, le probable.

Le traitement des nouvelles données numériques permet aussi d'accéder à de gigantesques volumes de traces sans signaux, notamment à travers les bases de données des grandes infrastructures de réseaux : transports, téléphone, électricité, etc. Agrégés et anonymisés, ces flux permettent de calculer de nouveaux services. Il est par exemple possible de prédire les files d'attente à Disneyland à partir du volume de communications sur le mobile depuis le RER A. Ces traces sans signaux produisent utilement des estimations statistiques grossières et globales. Mais voir davantage, c'est parfois ne rien voir. Dans le domaine de la surveillance, la capture massive des métadonnées du trafic de l'ensemble de la population semble vaine.

La prédiction est une probabilité qui comporte toujours une marge d'erreur créant des « faux positifs ». Si un algorithme était capable de détecter un comportement terroriste sur le réseau avec une marge d'erreur

de 1 %, ce qui constituerait déjà une prouesse, l'algorithme identifierait 600 000 personnes sur les 60 millions de Français. Si, en réalité, il n'y a que 60 terroristes, la surveillance des 599 940 autres semble totalement disproportionnée[14]. Sans doute est-il plus raisonnable d'obtenir de bons signaux au moyen de renseignements humains, pour ensuite se mettre à l'écoute des suspects[15].

La quantification de soi

Si les traces ne sont pas extraites des grandes infrastructures de réseaux, elles peuvent être produites par les individus eux-mêmes à travers les pratiques de mesure de soi (*quantified self*) et la diffusion des premiers objets communicants. Dans les mondes numériques, les petits groupes qui militent pour la quantification de soi soutiennent que la liberté expressive des individus doit aussi passer par son expression chiffrée. Prolongeant des pratiques anciennes, les capteurs numériques permettent aux internautes de mesurer, pour eux et pour les autres, une activité habituelle qui, devenue tableau statistique, devient aussi un miroir réfléchissant. Activités sportives, déplacements, temps de sommeil, signaux corporels, actes sexuels, les senseurs enregistrent les traces de comportements des personnes, pendant que les capteurs déposés dans l'environnement (voiture, compteur électrique, potager, pollution atmosphérique) mesurent son écosystème.

Deux orientations différentes peuvent être données à ces pratiques. Prolongeant la rationalisation réflexive des pratiques de soi, la première invite les individus à se doter d'outils d'auto-contrôle. La confrontation du sujet à la quantification de ses comportements est promue comme un instrument de construction de l'identité, un *benchmark* personnel. Les enquêtes menées auprès des premiers praticiens de ces outils font apparaître que leur utilisation est étroitement associée à un projet de perfectionnement de soi : optimiser son budget, maigrir, rester en forme, surveiller sa consommation ou son bilan carbone, etc. La relation que l'utilisateur entretient avec l'artefact constitue souvent un moyen de déléguer à un outil technique le soin de vérifier que l'on agit bien conformément à ses propres résolutions. Cependant, à l'exception des mesures sportives, il apparaît que

les utilisateurs abandonnent vite l'enregistrement de traces qui, sans être référées à d'autres signaux, sont souvent très monotones[16].

L'autre direction donnée à ces enregistrements personnels vise, dans une perspective fonctionnaliste, à créer un écosystème de mesures agrégeant différents flux, les partageant avec d'autres et les associant à des bases de données comportant des signaux plus riches. Le pèse-personne envoie des données vers Twitter, le rythme de la course enregistré par les chaussures détermine les choix musicaux de la *playlist* du jogger, les plantes réclament qu'on les arrose, les objets du quotidien partagent leurs données pour remplir le réfrigérateur, surveiller la maison ou ranger les affaires, etc.

De plus en plus, les objets qui nous environnent cherchent à désenclaver les mesures de leur enclos numérique pour se glisser dans les activités quotidiennes. Système fonctionnel, autostable, qui apprend constamment des rétroactions de l'utilisateur, ces nouveaux outils ouvrent d'immenses possibilités, mais posent en revanche de redoutables questions quant à la gouvernance de ces nouvelles données. La plupart du temps, les connaissances qu'ils peuvent apporter n'ont de pertinence que si, de l'utilisateur aux grandes bases de données, une ligne continue peut être activée. Soit les utilisateurs confient leurs enregistrements personnels à des services qui les croisent et les calculent, mais en perdant alors le contrôle sur leurs données personnelles ; soit le croisement entre les données est fait par l'utilisateur lui-même, qui deviendra alors l'administrateur de ses propres données, comme le proposent les initiatives de *self data* invitant les utilisateurs à devenir des collecteurs-interprètes de données. Dans tous les cas, le partage de bases de données publiques, libres et ouvertes réclame une gouvernance originale permettant de profiter des savoirs dont elles peuvent être le support, sans favoriser des phénomènes de centralisation et sans menacer la vie privée des personnes. Beaucoup de travaux ont montré qu'il suffit de très peu d'informations pour désanonymiser par recoupement des bases de données apparemment bien sécurisées.

Dans les enquêtes, les inquiétudes que suscite la capture des données personnelles n'ont jamais été aussi fortes, même si, à travers leurs conduites, les internautes ne semblent pas en tirer les conséquences. La vie privée n'a pas disparu et n'est pas devenue obsolète, comme l'annoncent en l'espérant les patrons des GAFA. En revanche, les conceptions de la vie privée

s'individualisent et cette « privatisation du privé » transforme les conditions dans lesquelles elle peut être protégée.

La vie privée a longtemps été pensée comme un bien collectif à partir duquel était érigé un ensemble de normes communes à tous, mais aussi de valeurs partagées par l'ensemble de la société comme le tact, la pudeur ou la discrétion. Cette définition univoque et générale de la vie privée est aujourd'hui fragilisée par le souci des individus d'en définir eux-mêmes la teneur et de ne pas laisser à d'autres le soin de décider de ses contours. Construite comme une protection, la vie privée est de plus en plus vécue comme une liberté. Les utilisateurs d'Internet souhaitent mieux contrôler ce qu'ils acceptent de rendre public ou de confier à d'autres.

Ce processus d'individualisation a favorisé l'idée que les utilisateurs procédaient à des arbitrages entre vie privée et sécurité (face à la surveillance étatique) ou entre vie privée et efficacité du service (face à l'instrumentalisation par le marché). Cette conception libérale et propriétaire des données personnelles laisse penser que les utilisateurs seraient en position de faire un choix libre et éclairé. Or tout montre que cette vision équilibrée et rationnelle de l'arbitrage est illusoire. Si les individus revendiquent le contrôle sur leurs données, ils le font dans un contexte où l'asymétrie des informations et l'absence d'alternative teintent leur choix de résignation[17].

À l'heure des *big data*, il est de moins en moins possible de connaître à l'avance le sens et la nature des calculs qui vont être conduits à partir des données collectées. La conception contractuelle d'une collecte des données finalisée et proportionnée, telle que la définissait la loi « Informatique et liberté » de 1978, a perdu son sens. Aussi le débat juridique sur la régulation des données personnelles s'oriente-t-il de plus en plus vers un renforcement du contrôle *ex post* de la régularité des traitements qui ont été réalisés par les opérateurs de données. Puisqu'il est difficile de choisir *ex ante* d'être ou de ne pas être tracé, il devient de plus en plus important de contrôler *ex post* ce dont nos données font l'objet. Il faut, pour cela, auditer les algorithmes.

Les algorithmes sont-ils biaisés ?

Le fonctionnement des algorithmes est un secret bien gardé[18]. Plus les individus sont transparents, plus ceux qui les observent sont opaques. Les grands acteurs du web protègent jalousement la propriété commerciale de la recette de leurs algorithmes, au prétexte que la rendre publique faciliterait la vie de ceux qui essaient de les truquer. Il est vrai que, pour certains grands calculateurs du web, celui de Google notamment, la connaissance détaillée des paramètres de son fonctionnement donnerait aux cohortes de sites en quête de visibilité l'occasion de prendre d'assaut la mesure, en dégradant pour tous la qualité des résultats.

Mais comment être sûr que les classements ne sont pas, en effet, truqués et trompeurs ? Pourquoi ce site est-il mieux classé que celui-là ? Pourquoi cette recommandation plutôt que telle autre ? Le profil personnel constitué à partir de mes traces ne peut-il conduire à des décisions aberrantes, injustes ou discriminatoires ? Les critiques qui doivent être adressées aux algorithmes sont différentes selon les familles de calcul que nous avons identifiées. Ils peuvent biaiser, tromper ou discriminer.

Les algorithmes sont d'abord accusés de déformer, trahir ou censurer la représentation « vraie » ou « neutre » de la réalité. Il ne peut cependant y avoir de « biais » que si l'on peut opposer aux classements algorithmiques une représentation bonne ou juste, comme le fait la statistique de la représentativité qui échantillonne sa population et donne un poids identique aux actions de chacun. Ce reproche n'a donc de sens que lorsque l'opération statistique produite par le calculateur est un simple dénombrement, comme dans les mesures produites à côté (audience) ou dans le web (réputation).

Le soupçon porte autant sur l'algorithme lui-même que sur ceux qui cherchent à le manipuler. Des robots cliqueurs augmentent artificiellement l'audience des sites ou le nombre de vues sur YouTube. Le marché des faux comptes Facebook ou Twitter permet aux entreprises en mal de e-réputation de gonfler artificiellement leurs compteurs et de se prévaloir d'une notoriété qu'elles ont en fait achetée. Hôteliers, commerçants et vendeurs de toutes sortes font appel à des petites mains pour rémunérer de faux avis de consommateurs (10 à 30 % des avis de consommateurs sur Internet sont des faux)[19]. Dans le monde de la e-réputation, où il est désormais admis d'agir stratégiquement pour construire sa visibilité, Facebook et Twitter ne font

guère d'efforts pour chasser les faux comptes et les tricheurs. Le site de notation Yelp a été soupçonné de faire écrire de mauvaises revues sur les produits des entreprises qui refusent de lui acheter de la publicité. Souvent, par complaisance ou par intérêt, les plateformes qui détiennent la mesure « ferment les yeux » sur les chiffres maquillés, les compteurs manipulés, les audiences élargies et les avis de faussaires. Les mondes numériques, qui étaient censés apporter de la précision là où les mesures du marketing traditionnelles étaient très floues, se sont résignés à accepter que les internautes entrent en compétition pour faire des mesures de réputation un thermomètre en caoutchouc.

Mais si un algorithme propose une opération plus complexe qu'une simple sommation, comme le font les mesures d'autorité, il n'existe pas de bonne « représentation » à partir de laquelle un « biais » puisse être mesuré. En se plaçant *au-dessus* du web, les calculateurs cherchent à rester les plus discrets possible afin de mesurer sans influencer. Le classement des sites de qualité par Google est une approximation réalisée à partir de règles et de paramètres très nombreux. Des centaines d'autres classements pourraient lui être opposés sans qu'un consensus sur la bonne représentation de l'information ne soit possible. Beaucoup de sites mal classés par Google ont essayé de faire des procès à l'entreprise américaine et ont systématiquement échoué devant sa ligne de défense systématique : l'algorithme fait mécaniquement un choix éditorial et, au nom de la liberté d'expression, il est libre de classer comme il le souhaite.

Il est pourtant difficile d'accepter cette défense, en raison du monopole tentaculaire qu'exerce aujourd'hui Google sur l'ensemble de l'écosystème numérique. Aussi, pour le critiquer, faut-il changer l'angle d'attaque. Si la « neutralité » des algorithmes est impossible à vérifier, il est en revanche important de demander aux plateformes du web de respecter leurs utilisateurs en faisant réellement faire à leurs calculateurs ce qu'elles disent et prétendent leur faire faire[20]. Dans de récents rapports, le Conseil national du numérique (CNNum) et le Conseil d'État ont fait apparaître la revendication nouvelle d'une *obligation de loyauté* des plateformes envers leurs utilisateurs[21].

L'obligation de loyauté interroge non pas une vaine objectivité ou vérité de la représentation des informations, mais l'alignement, ou le désalignement, entre le service que la plateforme prétend rendre et la réalité de ce qu'elle offre. Que Google privilégie ses propres services dans son classement (alors que ceux-ci ont moins d'« autorité » que d'autres), que Facebook donne une forte visibilité à certains contenus (alors que l'utilisateur n'a pas un fort « engagement » avec eux), qu'Amazon ajoute des livres à promouvoir dans ses recommandations (alors qu'ils ne correspondent pas à des utilisateurs ayant un profil d'achat similaire), et le service rendu par les algorithmes apparaîtra « déloyal ». Les algorithmes hiérarchisent les informations, et c'est pour cela qu'ils sont utiles et même essentiels. Mais il est indispensable que les services puissent expliquer à l'utilisateur les priorités qui président aux décisions de leurs calculateurs ; et qu'on puisse vérifier, en toute indépendance, que des intérêts cachés, des déformations clandestines ou des favoritismes n'altèrent pas le service rendu.

Lors du mouvement Occupy Wall Street aux États-Unis en 2011, certains militants se sont étonnés de ne pas voir les hashtags #occupy et #OWS (Occupy Wall Street) apparaître dans les « tendances » (*trending topics*) de Twitter. En volume, il ne fait guère de doute que le hashtag méritait d'apparaître dans cette fenêtre de visibilité. En revanche, sa vélocité était trop faible pour créer un pic, comme le font les annonces de décès de personnalités, les catastrophes naturelles ou les émissions de télévision à succès. Populaire, mais pas explosif, le calcul de la tendance n'était pas en faveur des militants américains, qui ont reproché à Twitter de censurer leur mouvement[22].

Les mesures fixent des règles pour calculer les événements qu'elles enregistrent. Elles leur donnent une forme. Twitter a décidé que les mouvements d'opinion devaient être immédiats et simultanés, associant ainsi la culture du direct télévisuel à celle de la viralité sur Internet. Comprendre et critiquer cette culture du « pic » attentionnel, c'est aussi encourager d'autres manières de mesurer et de donner de la visibilité à la circulation des opinions sur la Toile. Le développement d'une éducation et d'une culture partagées des algorithmes devrait nous aider à décoder et interpréter la manière dont ils façonnent nos représentations. À la suite de

protestations des mouvements féministes, Amazon a retiré en mai 2015 de l'interface de son site la possibilité de rechercher des jouets « garçons » ou des jouets « filles ». Sous la pression des utilisateurs, Facebook a dû revenir sur plusieurs de ses initiatives visant à introduire dans le fil d'actualité des informations qui n'étaient pas associées aux conversations des utilisateurs. C'est en soumettant les modèles à des audits indépendants qu'il est à la fois possible de prendre de la distance à l'égard du verdict des algorithmes et de leur opposer des calculs alternatifs.

L'« idiotie » des algorithmes

Les effets des algorithmes de la quatrième famille, ceux qui prédisent les comportements des internautes à partir des traces prélevées sous le web, sont beaucoup plus difficiles à critiquer. Leurs agissements se réalisent souterrainement, dans les bases de données des *data brockers*, et sans qu'il soit possible de comparer leurs résultats (puisqu'ils ont été personnalisés pour chacun). Parce qu'ils fonctionnent comme de purs automatismes procéduraux, les algorithmes donnent souvent des résultats statistiques imparfaits, stupides ou choquants.

Si, lorsque l'on tape le nom de certaines personnalités dans Google, le moteur de recherche suggère parfois d'y ajouter « juif », c'est parce que beaucoup d'internautes l'ont déjà fait. Lorsqu'au 1er janvier, Facebook propose aux utilisateurs un résumé illustré de leur année, l'algorithme sélectionne les publications qui ont suscité le plus d'interactions avec les amis, quitte à mettre en valeur la mort d'un proche. Les algorithmes suivent leurs procédures bêtement et ils manquent d'autant plus de tact et de sens moral que, ne calculant que des traces de comportements, ils font disparaître les catégories qui pourraient les empêcher de prendre en considération tel ou tel résultat. Les nouveaux calculateurs aspirent à être le reflet idiot d'une régularité statistique.

Latanya Sweeney, une chercheuse en informatique afro-américaine, a remarqué que, lorsqu'elle tapait son nom dans le moteur de recherche de Google, elle voyait apparaître la publicité « Latanya Sweeney, arrested ? ». Cette publicité propose un service de consultation en ligne,

instantcheckmate.com, qui permet, entre autres choses, de savoir si les personnes ont un casier judiciaire. Or le nom de ses collègues blancs n'était pas associé au même type de publicité, mais plutôt à des propositions commerciales de robes de mariée ou de retrouvailles avec les amis d'enfance[23].

Le système publicitaire de Google est-il discriminatoire ? Met-il en œuvre un fichier catégorisant les « Blancs » et les « Noirs » ? Après une enquête de rétro-ingénierie systématique sur un très grand nombre de requêtes, Latanya Sweeney a montré que l'algorithme n'a pas besoin d'avoir une intention classificatoire pour produire ce genre d'effets discriminatoires. Il ne comporte pas de règles lui demandant de détecter les personnes noires et les personnes blanches. Il se contente de laisser faire les régularités statistiques qui font que les noms et prénoms des personnes noires sont statistiquement plus souvent liés à des clics vers des recherches de casier judiciaire. Livré à lui-même, le calculateur s'appuie sur les comportements des autres internautes et contribue, « innocemment » si l'on ose dire, à la reproduction de la structure sociale, des inégalités et des discriminations.

L'individualisation des calculs, dans les grandes bases de données, produit des catégorisations sans en avoir l'air. Aux États-Unis, le « FICO score » mesure, pour chaque individu, les risques qu'il présente face au crédit à la consommation. Si ce fichier est public, beaucoup d'autres enregistrent et croisent dans la plus grande opacité des données concernant le profil des ménages, l'endettement, la consommation, la situation bancaire ou judiciaire. Sur cette question, les législations nationales sont plus ou moins tolérantes, mais partout le marché des données s'étend entre des entreprises qui se revendent ou s'échangent les informations. Cependant, en croisant des informations légales sur les individus, il est possible de faire des prédictions sur certains de leurs attributs, comme la sexualité, la religion ou les opinions politiques, dont il est interdit de faire des fichiers[24].

Non sans beaucoup d'approximations et d'erreurs, le croisement des informations personnelles permet de deviner de nouvelles informations. La fouille de données (*data mining*) dans les fichiers de relation-clients permet de ranger les profils en produisant discrètement des catégories construites à partir du seul intérêt du propriétaire des données : les clients intéressants

sont séparés de ceux qui ne le sont pas, les fidèles des infidèles, les « nouveaux prospects » sont convoités, etc. Ces pratiques anciennes sont en train d'étendre leur périmètre à travers l'interconnexion, par les entreprises, des bases de données de clientèle avec les traces qu'elles peuvent extraire des comportements sur le web.

Les techniques de *dynamic pricing* proposent de différencier les tarifs en fonction des profils, et l'on soupçonne certains services de pénaliser leurs clients les plus fidèles, sans alternatives ou pressés par le temps, en leur proposant des prix plus élevés. Tous ceux qui sont mal notés dans les bases de données, sans revenu, sans potentiel, endettés ou ayant un historique négatif, disparaissent du jeu des opportunités : ils n'accéderont pas à un bon taux de crédit, ne bénéficieront pas de coupons de réduction, ne recevront pas les informations, etc. De façon très conservatrice, le calcul algorithmique reconduit l'ordre social en ajoutant ses propres verdicts aux inégalités et aux discriminations de la société : les mal notés seront mal servis et leur note en deviendra plus mauvaise encore.

En mode automatique, les calculateurs n'ont plus besoin de catégories pour cibler les profils. En revanche, ceux qui manipulent les bases de données dans le monde du marketing ont toujours besoin de catégories pour donner du sens à leur activité et résister aux risques que l'automatisation fait peser sur leur métier. Disposant de données toujours plus nombreuses, les professionnels du *data mining* rangent les individus dans des segments de plus en plus fins sans jamais les en informer : « client pas fiable », « dépense médicale élevée », « revenu en déclin », « conduite à risque », « casanier avaricieux ». La méticuleuse précision des micro-segments favorise la multiplication de petites niches superposées qui découpent la société sans autre plan d'ensemble que celui d'agir efficacement et « commercialement » sur chacune d'elles.

Le service de vidéo de Netflix a ainsi créé près de 77 000 micro-genres pour classer les goûts de ses utilisateurs dans une suite de cases à la précision surréaliste[25] : « Drames sentimentaux européens des années 1970 avec paysages et couchers de soleil », « Comédie post-apocalyptique portant sur l'amitié », « Thrillers violents au sujet des chats pour les 8-10 ans », etc. Les catégories que produisent aujourd'hui les *big data* n'ont pas pour objectif d'être partagées avec les individus pour construire des catégories

d'identification offrant à la société un tableau d'ensemble. Elles découpent une interminable mosaïque de cibles pour préciser le tir des campagnes de marketing. Il n'est pas besoin que ceux qui sont identifiés dans la niche le sachent.

Il n'est d'ailleurs plus nécessaire de connaître les individus. L'ombre portée par la trace de leurs comportements, dans les fichiers informatiques, suffit à nourrir les calculs et à reconnaître les comportements similaires. Le calcul des traces n'a pas tellement l'individu ou le sujet pour cible[26]. Il n'est guère nécessaire que ceux qu'il identifie aient une psychologie, une histoire, une position sociale, des projets ou des désirs. Collection disparate de traces d'activités décousues révélant de façon kaléidoscopique des micro-facettes identitaires, l'individu calculé est un flux. Il est à la fois transparent et expulsé de ses propres traces.

Habituée à dénoncer l'hégémonie des médias traditionnels, la critique des algorithmes, ces nouveaux *gatekeepers*, leur reproche de censurer et de déformer les messages au nom des intérêts commerciaux ou de l'idéologie des firmes américaines qui les conçoivent. Cette critique n'est pas infondée, comme en atteste la pudibonderie d'Apple ou de Facebook, qui chassent toute nudité de leurs services. Il est cependant probable qu'elle ne prend pas à la racine le changement en cours dans le monde des algorithmes et reconduit un reproche ancien sur un monde nouveau.

En alignant leurs calculs personnalisés sur les comportements des internautes, les plateformes ajustent leurs intérêts économiques à la satisfaction de l'utilisateur. Sans doute est-ce à travers cette manière d'entériner l'ordre social en reconduisant les individus vers leurs comportements passés que le calcul algorithmique exerce sa domination. Il prétend leur donner les moyens de se gouverner eux-mêmes ; mais, réduits à leur seule conduite, les individus sont assignés à la reproduction automatique de la société et d'eux-mêmes. Le probable préempte le possible.

Paradoxalement, c'est au moment où les internautes s'attachent, par leurs représentations, leurs ambitions et leurs projets, à se penser comme des sujets autonomes et libérés des injonctions des prescripteurs traditionnels

que les calculs algorithmiques les rattrapent, par en dessous si l'on peut dire, en ajustant leurs désirs sur la régularité de leurs pratiques.

Notes

1. Valérie Peugeot, « L'ouverture des données publiques : convergence ou malentendu politique ? », *in* Bernard Stiegler (dir.), *Confiance, croyance, crédit dans les mondes industriels*, Paris, Éditions FYP, 2011.

2. Sandrine Cabut et Pascal Santi, « Insuffisance rénale. La parole est aux malades », *Le Monde*, 30 mars 2013.

3. Hubert Dreyfus, *What Computers Still Can't do. A Critique of Artificial Reason*, Cambridge, The MIT Press, 1992.

4. David Lazer, Ryan Kennedy, Gary King et Alessandro Vespignani, « The Parable of Google Flu : Traps in Big Data Analysis », *Science* n° 343, 14 mars 2014, p. 1203-1205.

5. Bilel Benbouzid, « De la prévention situationnelle au *predictive policing*. Sociologie d'une controverse ignorée », *Champ pénal*, vol. XII, juin 2015.

6. Eytan Bakshy, Solomon Messing et Lada Adamic, « Exposure to Ideogically Diverse News and Opinion on Facebook », *Science*, n° 348, 7 mai 2015.

7. Alex Pentland, *Social Physics. How Good Ideas Spread. The Lessons from a New Science*, New York, The Penguin press, 2014.

8. Katherine Milkman, Todd Rogers et Max Bazerman, « Highbrow Films Gather Dust : Time-Inconsistent Preferences and Online DVD Rentals », *Management Science*, vol. 55, n° 6, juin 2009, p. 1047-1059.

9. Gilles Deleuze, « Post-scriptum sur les sociétés de contrôle », in *Pourparlers. 1972-1990*, Paris, Minuit, 2003.

10. Dominique Boullier, *Les Sciences sociales face aux traces du big data ? Société, opinion et répliques*, FMSH-WP-2015-88, avril 2015.

11. Dominique Boullier et Audrey Lohard, *Opinion Mining et Sentiment Analysis*, Paris, OpenEdition Press, 2012.

12. Ian Ayres, *Super Crunchers*, *op. cit.*, chap. 6.

13. Nate Silver, *The Signal and the Noise. Why so Many Predictions Fail – but Some don't*, New York, Penguin Books, 2012.

14. Martin Untersinger, « La note interne de l'Inria qui étrille la loi sur le renseignement », *Le Monde*, 13 mai 2015.

15. Bruce Schneier, *Data and Goliath. The Hidden Battles to Collect Your Data and Control Your World*, New York, W. **W.** Norton & Company, 2014.

16. Anne-Sylvie Pharabod, « La mise en chiffre de soi. Une approche compréhensive des mesures personnelles », *Réseaux*, n° 177, 2013.

17. Joseph Turow, Michael Hennessy et Nora Draper, *The Tradeoff Fallacy. How Marketers are Misrepresenting American Consumers and Opening them up to Exploitation*, Philadelphie, University of Pennsylvania, 2015.

18. Franck Pasquale, *The Black Box Society. The Secret Algorithms That Control Money and Information*, Cambridge, Harvard University Press, 2015.

19. Joseph Reagle, *Reading the Comments. Likers, Haters, and Manipulators at the Bottom of the Web*, Cambridge, The MIT Press, 2015.

20. James Grimmelmann, « Speech Engines », *Minnesota Law Review*, n° 968, 2014.

21. CNNum, « Ambition numérique », rapport remis au Premier ministre, juin 2015 ; et Conseil d'État, « Rapport sur la neutralité des plateformes », Paris, La Documentation française, 13 juin 2014.

22. Tarleton Gillepsie, « Can an Algorithm be Wrong ? », *Limm*, n° 2, mars 2012.

23. Latanya Sweeney, « Google Ads, Black Names and White Names, Racial Discrimination, and Click Advertising », *ACM Queue*, n° 3, mars 2013.

24. Jernigan Carter, Mistree Behram, « Gaydar : Facebook Friendship Expose Sexual Orientations », *First Monday*, vol. 14, n° 10, 2009.

25. Alexis Madrigal, « How Netflix Reverse Engineered Hollywood », *The Atlantic*, 2 janvier 2014.

26. Antoinette Rouvroy et Thomas Berns, « Gouvernementalité algorithmique et perspective d'émancipation », *Réseaux*, n° 177, 2013.

CHAPITRE 4

La société des calculs

Il est important de porter un regard critique sur le fonctionnement des calculateurs plutôt que de les laisser agir silencieusement. Mais cet ouvrage voudrait aussi proposer une lecture plus politique du type de société qui rend aujourd'hui possible le déploiement des algorithmes.

La « tyrannie du centre »

La généralisation des calculs offre un point de vue original pour saisir, à l'état chiffré, la manière dont nos existences se transforment à l'heure des *big data*. Les algorithmes ont entrepris de calculer la société par le bas, depuis le comportement des internautes. Leurs concepteurs partagent l'idée que les informations ne doivent pas être choisies par les journalistes, que les publicités ne peuvent pas être les mêmes pour tous, que les catégories d'appartenance traditionnelles représentent mal les individus et que chacun doit pouvoir choisir librement ses « contenus » sans subir le paternalisme des prescripteurs. Les zélotes californiens des *big data* ont pour projet de refabriquer nos sociétés à partir d'un réel chiffré, plutôt que sur les fondements biaisés des idéologies, des intérêts et des programmes[1].

Les plus scientistes imaginent un monde enfin rationnel, débarrassé de l'encombrante subjectivité de ceux qui le gouvernent. Les autres promeuvent la vision libertarienne d'une société capable de s'auto-organiser et de sécréter les chiffres qui la représentent, en confiant au marché le soin de refléter ce que les États déforment. Les techniques de calcul mises en

œuvre pour réorganiser la société depuis les individus ont pris des formes multiples : les clics des internautes fabriquent de la popularité, les citations hypertextuelles de l'autorité, les échanges entre cercles affinitaires de la réputation, les traces des comportements une prédiction personnalisée et efficace.

Ces principes ne sont pas nouveaux. Ils partagent de nombreux traits avec les formes traditionnelles de représentation de la société. Simplement, leur manière d'asseoir leur légitimité sur des mesures prétendument objectives leur confère une force toute particulière. « Gangnam Style », la vidéo de Psy, chanteur coréen inconnu à qui personne n'aurait accordé la moindre chance, peut se prévaloir d'une popularité mondiale de 2,3 milliards de vues sur YouTube. Wikipédia, qui a successivement suscité l'ironie, l'hostilité puis le respect des élites lettrées, est le site à qui Google confère le plus d'autorité. Sur Twitter, YouTube et dans la blogosphère, une petite avant-garde d'influenceurs aux scores de réputation retentissants est apparu sans titre aucun pour fabriquer les tendances dans les secteurs de la mode, de l'humour ou de la cuisine[2]. Chaque internaute se voit proposer des recommandations personnalisées que chacun de ses nouveaux clics rend plus efficaces.

Le récit que nous avons proposé au premier chapitre a vu des principes de calcul établissant des normes collectives, la popularité et l'autorité, être de plus en plus concurrencées par des normes locales, la réputation, et des normes personnelles, la prédiction. Les calculateurs prétendent libérer la société de la « tyrannie du centre ». Pourtant, cette émancipation, par l'intermédiaire de mesures qui s'exercent sous la tutelle des intérêts économiques, continue de produire des effets de centralité d'autant plus forts qu'ils se sont largement émancipés des cadres nationaux pour devenir globaux. Le paradoxe de la société des calculs est qu'elle amplifie les phénomènes de coordination de l'attention et de hiérarchisation du mérite, tout en permettant aux individus de se sentir de plus en plus libres de leurs choix. En fait, les calculateurs donnent à la société les moyens de reproduire d'elle-même les inégalités et les hiérarchies qui l'habitent. Plus que jamais, il importe de savoir à quoi rêvent les algorithmes.

La coordination virale de l'attention

Les algorithmes rêvent d'un monde où les mécanismes de production de la popularité seraient transparents et ouverts à tous. Alors que le web est porteur de la double promesse d'un élargissement de l'offre d'informations et d'une distribution plus étale des consommations (phénomène appelé « longue traîne »), tout montre qu'on assiste à une sur-concentration de l'attention autour de certaines informations qui gagnent une immense, soudaine et brève popularité en raison des effets de coordination virale qui orientent les publics vers quelques produits[3].

La présence insistante des compteurs de « nombre de vues », tout au long du parcours de navigation des internautes, contribue à produire ces pics d'attention. Les traditionnelles mesures d'audience continuent de fidéliser les publics. Dans le domaine des médias, de la politique ou de la culture, la hiérarchie des popularités n'a pas été radicalement renversée par Internet. Mais la fabrication de la popularité numérique est désormais versatile, brusque et déroutante. Elle privilégie la synchronisation, le mimétisme et l'obsolescence programmée. La consécration des célébrités, au terme d'une trajectoire ritualisée ponctuée de seuils et d'épreuves, est bousculée. Des humoristes en chambre aux chaînes de beauté sur YouTube, des vidéos de sensibilisation à des causes humanitaires aux blagues à effets viraux, les mondes numériques sont traversés de popularités improbables et flottantes, qui marquent des écarts différentiels extrêmement forts entre des produits aux qualités *a priori* très similaires.

Un appareillage de supervision et de mesure s'est introduit de plus en plus profondément dans la conception des informations et des messages. Au sein des rédactions, les journalistes sont partagés sur la nécessité d'accéder à des outils de monitoring en temps réel de l'audience de leurs articles. Mais, dans les nouveaux médias en ligne en quête de *buzz*, l'information est soumise à une loi de « sélection naturelle » : le même article est publié sous trois titres différents pendant 30 minutes, avant que ne soit conservé celui qui capitalise la plus forte audience[4]. La production d'informations « à cliquer » (*clickbait*) jouant sur des ressorts émotionnels, émoustillants ou comiques engendre une industrie nouvelle et répand des formats d'écriture

(« Les 10 raisons de… », « Découvrez comment… ») auxquels les médias traditionnels ont de plus en plus de mal à résister.

Par un effet d'autorenforcement alimenté par les compteurs d'audience, l'attention attire l'attention. Elle produit de nouvelles formes de célébrité dont il est difficile d'identifier les mérites. Lorsque l'offre d'information abonde, c'est désormais l'attention des publics qui constitue un bien rare et convoité[5].

Il faut cependant apporter deux nuances à l'idée trop répandue qu'Internet serait la cour de récréation des rumeurs, des fausses valeurs et des engouements éphémères. Il est assez rare, dans les classements, que les célébrités de YouTube dominent celles des industries culturelles. Ces dernières savent très bien donner une existence virale à leurs créatures et les techniques de marketing de masse continuent de produire, à échelle mondiale, les principaux foyers d'attraction. Par ailleurs, les comportements des internautes ne se réduisent pas à des comportements automatiques et des séductions immédiates. Alors que les tenants de la confrontation entre anciens et nouveaux médias cherchent à séparer les internautes « automatisés » – les jeunes, les extrémistes, les anonymes – des lecteurs réfléchis et vigilants, il est beaucoup plus raisonnable de considérer que c'est en fait la gamme des formes d'attention qui s'est ouverte pour tous[6].

Selon les contextes, elle ménage des places pour l'absorption envoûtée, la dispersion flottante, le grégarisme automatique, comme pour l'exploration approfondie, le vagabondage vers les périphéries ou l'investigation critique. Sans doute l'enjeu est-il plutôt de savoir quel soin nous prenons pour faire de notre (rare) attention le meilleur usage possible parmi les multiples sollicitations des industries de la popularité[7].

La sécession des excellents

Avec les mesures d'autorité, les algorithmes rêvent d'un monde où la reconnaissance des « méritants » ne serait pas entravée : ils veulent désigner les excellents et valoriser les meilleurs. Le monde vu par Google est un univers méritocratique qui confère une visibilité disproportionnée aux sites

web les plus reconnus. Il fait de la planète numérique le terrain d'une gigantesque compétition pour l'excellence et réserve aux élus un minuscule nombre de places.

Souvent critiquée comme une évaluation imparfaite et falsifiable de la qualité, cette technique de calcul entretient une forte proximité avec les valeurs méritocratiques. Refusant les positions héritées et statutaires, elle agrège la notoriété en évaluant les agissements indépendamment des places occupées dans la société. Les métriques d'autorité prétendent donner à chacun l'occasion de faire reconnaître ses qualités à travers ses accomplissements. Mais, ce faisant, elles font disparaître la structure des places permettant de brider la concurrence pour les talents qui, lorsqu'elle s'appuie sur une mesure de reconnaissance, exacerbe les inégalités[8].

La fluidification et la globalisation du marché des jugements numériques engendrent des effets statistiques qui confèrent aux gagnants des avantages cumulés considérables[9]. Il est frappant de constater que les traditionnelles distributions des inégalités selon la loi de Pareto, qui donne à 20 % d'une population 80 % du bien à répartir, ont pris sur le web la forme d'une loi de puissance beaucoup plus accusée, qui réserve souvent à moins de 1 % des acteurs plus de 90 % de la visibilité. Ce glissement vers des répartitions extrêmes est une conséquence des calculs en réseau.

« L'effet Matthieu » est une loi de distribution des avantages cumulés qui, dans les structures en réseau, favorise ceux qui occupent une position centrale. L'augmentation des inégalités dans nos sociétés se fait moins autour des valeurs moyennes qu'à travers une accélération, au sommet de la pyramide, de l'accumulation par les mieux dotés d'une part considérable de la ressource à distribuer. C'est de plus en plus le cas pour la richesse, pour la visibilité ou pour la notoriété. Chaque année, le Kunstkompass classe la liste des 100 artistes contemporains vivants les plus réputés, en mettant en œuvre un calcul d'autorité qui a tout du PageRank de Google : il donne des points aux artistes selon les expositions qu'ils ont faites dans tel ou tel musée, les musées disposant eux-mêmes d'une autorité différente en fonction du nombre d'artistes bien classés dont ils ont fait la promotion[10].

À l'instar des mondes de l'art qui se rêvent depuis longtemps comme un marché de concurrence pure et parfaite où les inégalités de succès sont

acceptées et célébrées, la mesure d'autorité propose une société où les talentueux raflent systématiquement la mise dans le journalisme, la mode, l'édition, le design, les métiers de service, le management des entreprises ou le monde universitaire[11]. Alors que les hiérarchies locales, thématiques et contextuelles permettent de maintenir un monde composite et pluriel, l'unification des marchés du jugement contribue à donner aux « meilleurs » une visibilité surnuméraire.

Classements de grandes écoles, cours en ligne assurés par les meilleures universités, notations financières, constitution de pôles d'excellence, palmarès international des systèmes scolaires ou médicaux, ces nouvelles mesures d'autorité s'émancipent des contextes dans lesquels les activités étaient traditionnellement mesurées avant d'être comparées à d'autres. L'autorité des excellents fabrique « des gagnants individualisés et des perdants invisibilisés[12] » en demandant aux perdants de produire les signes de reconnaissance qui donnent aux gagnants l'illusion d'être propriétaires de leurs qualités. Dès lors, la redistribution des positions sociales par la reconnaissance du mérite permet aux excellents de faire sécession. Dans ce grand écart, nos sociétés sont en train d'oublier la moyenne.

Digital labor

À travers les métriques de réputation du web social, les algorithmes rêvent aussi à une société dans laquelle ils donneraient aux personnes des outils pour que les affinités puissent se reconnaître et s'auto-organiser. L'émancipation de la société des catégories institutionnelles qui permettaient de l'articuler a facilité le déploiement, par le bas, de réseaux affinitaires dans lesquels les individus expriment leurs singularités sans se plier aux assignations de rôle ou de statut. La crise de confiance à l'égard de la politique, des institutions, des journalistes ou des experts ne s'exerce pas à l'endroit des liens affinitaires, proches, locaux ou spécialisés. Sans doute sont-ils même devenus le refuge de beaucoup d'espérances autrefois adressées aux instances de représentation sociale et politique[13].

La valorisation du réseau relationnel, cet espace de confiance dans lequel il est possible de se réunir selon les engagements, les amitiés, les goûts et les idées partagés, apparaît dans toutes les enquêtes comme un vecteur de socialisation qui ne subit pas la crise de confiance que traversent nos sociétés. L'expressivité du web social a mis en visibilité la multiplication de ces nouveaux agrégats de « liens faibles », collectifs parfois fragiles et furtifs, parfois experts et durables, que sont les forums de jeux vidéo, les rédacteurs bénévoles d'encyclopédie en ligne, les collectionneurs de toutes sortes, les cartographes amateurs d'Open Street Map ou les participants des nouvelles plateformes de partage et d'échange (d'appartements, de voitures, etc.)[14].

En redéployant les identités et les manières de les associer, le web a permis la fabrication d'un ensemble inattendu de formes d'actions collectives : pétitions en ligne, mouvement de solidarisation pour une cause autour d'une page Facebook, déclenchement « spontané » de mouvements sociaux, financement coopératif de projets (*crowdfunding*), etc. Les univers que s'ouvrent ainsi les individus se caractérisent par le fait qu'ils ne préjugent pas *a priori* des identités des participants. C'est dans la constitution des coopérations et des fabrications collectives que, de façon interactive, l'identité des personnes, leur personnalité dans le réseau, leurs compétences et leurs qualités se construisent d'une manière qui n'est jamais figée.

On comprend alors comment les métriques de réputation, ces micro-repères qui calculent la force des engagements, les rapprochements entre goûts ou les états de service au sein de la communauté, ont pu trouver des moyens de se déployer dans ces nouveaux univers affinitaires. La recomposition de la société à partir des investissements expressifs des individus constitue sans doute la part la plus positive des nouvelles formes de vie numérique.

Cependant, avec le web des réputations, la visibilité s'est aussi imposée comme un nouveau principe de hiérarchisation de la valeur sociale. Le calcul d'autorité que réalise Google se doit de séparer strictement les actions naturelles et les actions stratégiques pour ne pas abîmer sa mesure. Le modèle économique du géant américain reconduit la séparation

traditionnelle entre éditorial et publicité, en considérant que la visibilité soit se mérite, soit s'achète.

Ce découpage entre l'authentique et le calcul n'a pas résisté aux usages tumultueux du web social. Les internautes ont multiplié les comportements destinés à « jouer » avec les métriques pour obtenir une meilleure visibilité. Le calcul de l'identité numérique valorisant la réputation dans le réseau instaure une distance entre la personne et son avatar, qui introduit une manière de se percevoir sous le mode de la construction et de la fabrication. La frontière entre don et calcul apparaît plus indéterminée. En témoigne la remise en cause de plus en plus fréquente de la sacro-sainte frontière entre éditorial et publicité dans le monde des médias en ligne. Par tous les moyens, les annonceurs essaient de glisser subrepticement leurs messages dans les flux des internautes, en laissant entendre à leurs clients qu'ils forment avec eux une « communauté ».

Sur Facebook, les marques parlent à des « amis » en les tutoyant. Couverts de cadeaux par les entreprises, les blogueurs réputés sont parfois soupçonnés d'écrire des billets de complaisance. Le *native advertising* propose des formats publi-rédactionnels dans lesquels les marques se cachent derrière des contenus attrayants et personnalisés. La justification de ces pratiques est que, si les utilisateurs revendiquent de choisir eux-mêmes les contenus qui les intéressent, s'ils ont parfois pour eux-mêmes des stratégies d'autopromotion, ils sont aussi capables de détecter les stratégies publicitaires et de les refuser.

Éditeurs, publicitaires et internautes, tous ensemble réflexifs et calculateurs, fabriquent un monde qui ne croit guère à l'authenticité. Le brouillage de cette frontière contribue aussi à regarder les expressions gratuites et passionnées des internautes comme un travail méritant une rémunération[15]. L'expressivité créative des internautes, les productions de toutes sortes qu'ils échangent, en suscitant de l'attention et de la renommée, sont capturées par les plateformes pour être monétisées. Après tout, soutiennent les tenants du *digital labor*, en entérinant la suppression de la frontière entre don et calcul, il serait logique qu'ils perçoivent une rémunération en récompense de la valeur produite par leur travail[16].

Pourtant, tous les individus ne disposent pas des mêmes ressources sociales et culturelles pour profiter des espaces de valorisation de soi. La « bulle » dans laquelle Facebook enferme ses utilisateurs censure moins des contenus selon les convictions des utilisateurs qu'elle n'oppose les « individus par excès »[17], « individus individualisés » qui font feu de tout bois pour faire briller leur réputation, et les « individus par défaut », qui se trouvent relégués à distance des informations d'actualité.

Dans une enquête permettant de dénombrer les liens partagés sur leur page Facebook, on a pu montrer que, sur une période de deux ans, 41 % des comptes Facebook n'avaient pas partagé un seul lien vers des médias d'information et 15 % un seul lien. En revanche, ces utilisateurs échangent dans leurs très vivantes conversations des liens nombreux vers des sites d'humour, de divertissement et des vidéos sur YouTube[18]. Il n'est pas besoin que Facebook biaise ou censure son algorithme ; il suffit qu'il le laisse calculer l'information que cliquent ses utilisateurs en reproduisant l'inégale répartition du capital culturel selon les réseaux relationnels[19].

En épousant les comportements des internautes, l'algorithme de Facebook reconduit les inégalités de nos sociétés en donnant aux mieux dotés les moyens d'enrichir leurs réseaux relationnels et d'accéder à plus de ressources et d'opportunités. Il est à craindre que, pour les autres, l'ouverture informationnelle ne produise pas les mêmes effets. À la différence des métriques d'audience ou d'autorité, les compteurs du web social ne sont pas faits pour aider l'Internet expressif à rencontrer l'Internet silencieux. Ils laissent chacun construire son espace informationnel en fonction de ses ressources sociales et culturelles.

« Passer en manuel »

Sans doute le rêve ultime des nouveaux calculs est-il d'installer un environnement technique invisible permettant partout et pour tout de nous orienter sans nous contraindre. Amatrice de science-fiction, la critique des algorithmes dramatise à l'envi les risques totalitaires d'une rationalisation des existences. La diversité des informations, l'exposition à des

connaissances multiples, la variété des choix ont connu une explosion si massive avec Internet qu'il est toujours surprenant d'entendre certains soutenir que nous serions enfermés dans la prison des algorithmes et des plateformes. Non sans contradiction, ce sont souvent les mêmes qui s'indignent qu'Internet permette aux idées extrémistes, aux thèses conspirationnistes et aux prêches fondamentalistes de rencontrer un public sur le réseau. Cette crainte révèle plutôt les contradictions que rencontre l'expansion du projet d'autonomie et de liberté individuelle face à la multiplication des opportunités de réalisation de soi[20].

Avec l'augmentation du capital culturel, plus les acteurs sociaux sont encastrés dans des univers multiples et interdépendants, plus ils chérissent l'idée d'une autodétermination du sujet et s'inquiètent des contraintes qui, de l'extérieur, pourraient s'exercer sur leur libre arbitre. Cette manière de tenir l'individu à distance des machines est très mal adaptée à l'analyse des nouvelles formes de sujétion qu'installe la société des calculs. Les infrastructures des *big data* cherchent à guider sans contraindre, à orienter sans obliger. Elles constituent un exemple typique de ce que Cass Sunstein appelle des *nudges*, ces outils du « paternalisme libertaire[21] » qui, par défaut, suppléent les choix des individus en les persuadant qu'ils agissent au mieux de leurs intérêts. En fait, les algorithmes rêvent de délester les humains de ce qu'il y a de plus mécanique dans leurs activités, assurant qu'ils les libèrent pour des tâches cognitives plus hautes, plus complexes ou plus ambitieuses.

La vie quotidienne des milieux sociaux les plus aisés est constamment sous-tendue par des choix qui, « par défaut », sont délégués à des agents de services et des infrastructures sociotechniques. Une quantité considérable de choix pratiques sont faits par d'autres à la place des personnes qui bénéficient du confort de ne pas avoir à se consacrer à mille et une décisions pratiques. Le *choix de ne pas choisir* est socialement distribué, et il constitue un luxe et une ressource qui ont été produits par les grandes architectures sociotechniques de nos sociétés. L'imaginaire des concepteurs de services de *big data* est obnubilé par la figure du concierge ou de l'assistant personnel.

Dans un monde de magazine pour cadres suractifs, les nouveaux calculs optimisent leur temps, prennent les billets d'avion, traduisent automatiquement, détectent le meilleur restaurant, répondent aux mails d'informations pratiques, trouvent avec qui sortir et remplissent le réfrigérateur, etc. Ceux qui s'inquiètent de la perte des savoir-faire humains sont souvent ceux qui, dans leur vie quotidienne et souvent sans s'en rendre compte, ont le plus finement ajusté leur vie aux trajectoires automatisées et confortables du guidage par les infrastructures.

Il n'en reste pas moins que l'enjeu politique que posent les nouvelles boîtes noires du calcul algorithmique est celui de la capacité à les débrayer et à « passer en manuel ». Le risque que présentent les nouvelles infrastructures de calcul est d'architecturer les choix en les fermant sur des processus irréversibles. Les calculateurs se proposent d'automatiser ce que nos vies comportent de mécanique, de fonctionnel et de statistique. Dans les activités complexes, les habiletés manuelles ont été transférées vers les machines. Les pilotes d'avion ne conduisent plus vraiment les avions, mais les surveillent. Les architectes ne font plus de dessins à la main, mais modélisent directement en 3D. Les algorithmes de détection visuelle sont en train d'apprendre à lire les radiographies et les IRM que valideront ensuite les médecins[22]. Face à ces grands systèmes techniques qui capturent nos habiletés, il est de plus en plus nécessaire d'apprendre à ne pas désapprendre[23].

Regarder la société depuis les calculateurs donne l'idée trompeuse que les internautes se plient aux desiderata des algorithmes. Les rêves des algorithmes ne sont que des rêves. Lorsque les usages sont observés depuis la réalité quotidienne des internautes, l'emprise des calculateurs sur leur vie semble s'évaporer. Les usages sont beaucoup plus vagabonds, diversifiés et stratèges que ne le pensent ceux qui raisonnent depuis une seule plateforme. Dans les enquêtes, les utilisateurs trouvent la personnalisation publicitaire médiocre, redondante et, la plupart du temps, à côté de la plaque. Elle ne parvient à attirer qu'un nombre infime de clics.

Il est des moments où les internautes choisissent le confort du guidage et d'autres où ils débrayent pour explorer et se perdre. La réduction de leurs pratiques à des automatismes comportementaux fait oublier que les usages

d'Internet ne cessent de se complexifier, de s'intellectualiser et de devenir en eux-mêmes des objets réflexifs. Comme pour toutes les autres technologies intellectuelles, il n'y a pas de raison de penser que les utilisateurs ne parviennent pas à socialiser les calculateurs, à déployer des stratégies pour les domestiquer et à leur opposer des contre-calculs, comme le montrent déjà les collectifs d'appropriation citoyenne des mesures de pollution et les initiatives qui se multiplient pour auditer les algorithmes.

Plutôt que de dramatiser le conflit entre les humains et les machines, il est plus judicieux de les considérer comme un couple qui ne cesse de rétroagir et de s'influencer mutuellement. La société des calculs réalise un couplage nouveau entre une puissance d'agir de plus en plus forte des individus et des systèmes sociotechniques imposant, eux aussi, des architectures de plus en plus fortes. Il est encore temps de dire aux algorithmes que nous ne sommes pas la somme imprécise et incomplète de nos comportements.

Notes

1. Tim O'Reilly, « Open Data and Algorithmic Regulation », *in* Brett Goldstein et Lauren Dyson (dir.), *Beyond Transparency. Open Data and the Future of Civic Innovation*, San Francisco, Code for America, 2013.

2. Dominique Cardon, Guilhem Fouetillou et Camille Roth, « Topographie de la renommée en ligne : un modèle structurel des communautés thématiques du web français et allemand », *Réseaux*, n° 188, 2014.

3. Kevin Mellet, « Aux sources du marketing viral », *Réseaux*, n° 157-158, 2009.

4. Bill Wasik, *And Then There This. How Stories Live and Die in Viral Culture*, New York, Viking Books, 2009.

5. Emmanuel Kessous, Kevin Mellet et Mustapha Zouinar, « L'économie de l'attention : entre protection des ressources cognitives et extraction de la valeur », *Sociologie du travail*, n° 52, 2010.

6. Dominique Boullier, « Composition médiatique d'un monde commun à partir d'un pluralisme des régimes d'attention », *in* Pierre-Antoine Chardel (dir.), *Conflit des interprétations dans la société de l'information*, Paris, Hermès, 2012.

7. Yves Citton, *Pour une écologie de l'attention*, Paris, Seuil, 2014.

8. François Dubet, *Les Places et les Chances. Repenser la justice sociale*, Paris, Seuil/La République des Idées, 2010.

9. Robert Franck et Philippe Cook, *The Winner-Take-All Society*, New York, The Free Press, 1995.

10. Alain Quemin, *Les Stars de l'art contemporain. Notoriété et consécration artistiques dans les arts visuels*, Paris, CNRS Éditions, 2013.

11. Pierre-Michel Menger, *Portrait de l'artiste en travailleur. Métamorphose du capitalisme*, Paris, Seuil/La République des Idées, 2002.

12. Fabienne Brugère, *La Politique de l'individu*, Paris, Seuil/La République des Idées, 2013, p. 13.

13. Monique Dagnaud, *Génération Y. Les jeunes et les réseaux sociaux, de la dérision à la subversion*, Paris, Les Presses de Science Po, 2011.

14. Patrice Flichy, *Le Sacre de l'amateur. Sociologie des passions ordinaires*, Paris, Seuil/La République des Idées, 2010.

15. Jaron Lanier, *Who Owns the Future*, San Jose, Simon & Shuster, 2013.

16. Dominique Cardon et Antonio Casilli, *Qu'est-ce que le* digital labor ?, Paris, INA Éditions, 2015.

17. Robert Castel, *La Montée des incertitudes. Travail, protection, statuts de l'individu*, Paris, Seuil, 2009.

18. Les résultats du projet ANR « ALGOPOL » sont accessibles sur *http://algopol.huma-num.fr*

19. François Héran, « La sociabilité, une pratique culturelle », *Économie et Statistique*, n° 216, 1998.

20. Axel Honneth, *Le Droit de la liberté. Esquisse d'une éthicité démocratique*, Paris, Gallimard, 2015.

21. Cass Sunstein, *Why Nudge ? The Politics of Libertarian Paternalism*, New Haven, Yale University Press, 2014.

22. Nicholas Carr, *The Glass Cage. Automation and Is*, New York, W. W. Norton & Company, 2015.

23. Bernard Stiegler, *La Société automatique. 1. L'avenir du travail*, Paris, Fayard, 2015.

CONCLUSION

La route et le paysage

Les médias ont longtemps été pour nous des gyrophares. Ils nous prenaient par la main pour nous mener tout en haut de la montagne. Là, depuis la table d'orientation, ils nous désignaient les éléments notoires du paysage : le sacro-saint « panorama ». Placée au centre de la société, la table d'orientation organise l'attention collective afin que tous, prêtant attention aux mêmes motifs, partagent le sentiment de faire société commune. Avec Internet, l'orientation de notre attention est libérée au motif que chacun, depuis son propre véhicule, peut se déplacer librement dans un univers proliférant d'informations. Contre le paternalisme de la table d'orientation et son panorama obligé, chacun peut organiser son voyage sans avoir à suivre les directions prescrites par un guide, fût-il de haute montagne.

La conquête de cette liberté est sans pareille. Elle déboussole, fait perdre du temps et prendre des risques. Elle ouvre à chacun le droit d'errer, de se tromper et de s'émerveiller. Pourtant, perdu parmi les mille et un choix possibles, il a fallu trouver d'autres manières de se repérer et de s'organiser. Les calculateurs ont apporté une solution originale et audacieuse à cette désorientation. Ils constituent un moyen efficace pour trier dans l'abondance d'informations disponibles et pour guider l'utilisateur vers ses propres choix.

Comme les GPS dans les véhicules, les algorithmes se sont glissés silencieusement dans nos vies. Ils ne nous imposent pas la destination. Ils ne choisissent pas ce qui nous intéresse. Nous leur donnons la destination et ils nous demandent de suivre « leur » route. La conduite sous GPS s'est si fortement inscrite dans les pratiques des conducteurs que ceux-ci ont parfois

perdu toute idée de la carte, des manières de la lire, de la diversité de ses chemins de traverse et des joies de l'égarement.

Les algorithmes nous ont libérés des voyages de groupe, des points de vue obligés et des arrêts obligatoires devant des panoramas à souvenirs. Ils procèdent d'un désir d'autonomie et de liberté. Mais ils contribuent aussi à assujettir l'internaute à cette route calculée, efficace, automatique, qui s'adapte à nos désirs en se réglant secrètement sur le trafic des autres. Avec la carte, nous avons perdu le paysage. Le chemin que nous suivons est le « meilleur » pour nous.

Mais nous ne savons plus bien identifier ce qu'il représente par rapport aux autres trajets possibles, aux routes alternatives et peu empruntées, à la manière dont la carte compose un ensemble. Nous n'allons pas en revenir aux voyages de groupe et à leur guide omniscient. En revanche, nous devons nous méfier du guidage automatique. Nous pouvons le comprendre et soumettre ceux qui le conçoivent à une critique vigilante. Il faut demander aux algorithmes de nous montrer et la route, et le paysage.

Dans la même collection

Éric MAURIN

L'Égalité des possibles
(2002)

Thérèse DELPECH

Politique du chaos
(2002)

Olivier ROY

Les Illusions du 11 septembre
(2002)

Jean-Paul FITOUSSI

La Règle et le Choix
(2002)

Michael IGNATIEFF

Kaboul-Sarajevo
(2002)

Daniel LINDENBERG

Le Rappel à l'ordre

(2002)

Pierre-Michel MENGER

Portrait de l'artiste en travailleur
(2003)

Hugues LAGRANGE

Demandes de sécurité
(2003)

Xavier GAULLIER

Le Temps des retraites
(2003)

Suzanne BERGER

Notre première mondialisation
(2003)

Robert CASTEL

L'Insécurité sociale
(2003)

Bruno Tertrais

La Guerre sans fin
(2004)

Thierry Pech, Marc-Olivier Padis

Les Multinationales du cœur
(2004)

Pascal Lamy

La Démocratie-monde
(2004)

Philippe Askenazy

Les Désordres du travail
(2004)

François Dubet

L'École des chances
(2004)

Éric Maurin

Le Ghetto français
(2004)

Julie Allard, Antoine Garapon

Les Juges dans la mondialisation
(2005)

François Dupuy

La Fatigue des élites
(2005)

Patrick Weil

La République et sa diversité
(2005)

Jean Peyrelevade

Le Capitalisme total
(2005)

Patrick Haenni

L'Islam de marché
(2005)

Marie DURU-BELLAT

L'Inflation scolaire
(2006)

Jean-Louis MISSIKA

La Fin de la télévision
(2006)

Daniel COHEN

Trois Leçons sur la société post-industrielle
(2006)

Louis CHAUVEL

Les Classes moyennes à la dérive
(2006)

François HÉRAN

Le Temps des immigrés
(2007)

Dominique MÉDA, Hélène PÉRIVIER

Le Deuxième Âge de l'émancipation
(2007)

Thomas PHILIPPON

Le Capitalisme d'héritiers
(2007)

Youssef COURBAGE, Emmanuel TODD

Le Rendez-vous des civilisations
(2007)

Robert CASTEL

La Discrimination négative
(2007)

Laurent DAVEZIES

La République et ses territoires
(2008)

Gösta Esping ANDERSEN

(avec Bruno Palier)

Trois Leçons sur l'État-providence
(2008)

Loïc BLONDIAUX

Le Nouvel Esprit de la démocratie
(2008)

Jean-Paul FITOUSSI, Éloi LAURENT

La Nouvelle Écologie politique
(2008)

Christian BAUDELOT, Roger ESTABLET

L'Élitisme républicain
(2009)

Éric MAURIN

La Peur du déclassement
(2009)

Patrick PERETTI-WATTEL, Jean-Paul MOATTI

Le Principe de prévention
(2009)

Esther DUFLO

Le Développement humain
Lutter contre la pauvreté (I)
(2010)

Esther DUFLO

La Politique de l'autonomie
Lutter contre la pauvreté (II)
(2010)

François DUBET

Les Places et les Chances
Repenser la justice sociale
(2010)

Dominique CARDON

La Démocratie Internet
Promesses et limites
(2010)

Dominique BOURG, Kerry WHITESIDE

Vers une démocratie écologique
Le citoyen, le savant et le politique
(2010)

Patrice FLICHY

Le Sacre de l'amateur
(2010)

Camille LANDAIS, Thomas PIKETTY, Emmanuel SAEZ

Pour une révolution fiscale

Un impôt sur le revenu pour le XXI^{e} siècle
(2011)

Pierre LASCOUMES

Une démocratie corruptible
(2011)

Philippe AGHION, Alexandra ROULET

Repenser l'État
Pour une social-démocratie de l'innovation
(2011)

Collectif

Refaire société
(2011)

Dominique GOUX, Éric MAURIN

Les Nouvelles Classes moyennes
(2012)

Blanche SEGRESTIN, Armand HATCHUEL

Refonder l'entreprise
(2012)

Nicolas DUVOUX

Le Nouvel Âge de la solidarité
Pauvreté, précarité et politiques publiques
(2012)

François BOURGUIGNON

La Mondialisation de l'inégalité

(2012)

Laurent DAVEZIES

La crise qui vient
La nouvelle fracture territoriale
(2012)

Michel KOKOREFF, Didier LAPEYRONNIE

Refaire la cité
L'avenir des banlieues
(2013)

Hervé LE BRAS, Emmanuel TODD

Le Mystère français
(2013)

Camille PEUGNY

Le Destin au berceau
Inégalités et reproduction sociale
(2013)

Fabienne BRUGÈRE

La Politique de l'individu
(2013)

Gabriel ZUCMAN

La Richesse cachée des nations
(2013)

Marie DURU-BELLAT

Pour une planète équitable
L'urgence d'une justice globale
(2014)

Antoine VAUCHEZ

Démocratiser l'Europe
(2014)

François DUBET

La Préférence pour l'inégalité
Comprendre la crise des solidarités
(2014)

Claudia SENIK

L'Économie du bonheur
(2014)

Julia CAGÉ

Sauver les médias
Capitalisme, financement participatif et démocratie
(2015)

Laurent DAVEZIES

Le Nouvel Égoïsme territorial
Le grand malaise des nations
(2015)

Éric MAURIN

La Fabrique du conformisme
(2015)

Table des matières

Google, Facebook, Amazon, mais aussi les banques et les assureurs : la constitution d'énormes bases de données (les « *big data* ») confère une place de plus en plus centrale aux algorithmes. L'ambition de ce livre est de montrer comment ces nouvelles techniques de calcul bouleversent notre société. À travers le classement de l'information, la personnalisation publicitaire, la recommandation de produits, le ciblage des comportements ou l'orientation des déplacements, les méga-calculateurs sont en train de s'immiscer, de plus en plus intimement, dans la vie des individus. Or, loin d'être de simples outils techniques, les algorithmes véhiculent un projet politique. Comprendre leur logique, les valeurs et le type de société qu'ils promeuvent, c'est donner aux internautes les moyens de reprendre du pouvoir dans la société des calculs.

Dominique Cardon est sociologue au Laboratoire des usages d'Orange Labs et professeur associé à l'université de Marne-la-Vallée (LATTS). Avec *La Démocratie Internet* (Seuil/La République des Idées, 2010) et de nombreux articles, il s'est imposé comme l'un des meilleurs spécialistes du numérique et d'Internet.

www.seuil.com et www.repid.com

9 782021 279962

ISBN 978.2.02.127996.2/Imprimé en France 10.2015 11,80 €