Enjeux numériques

Big Data :
économie et régulation

UNE SÉRIE DES
ANNALES
DES MINES
FONDEES EN 1794

N° 2 - Juin 2018

Publié avec le soutien
de l'Institut MinesTélécom

UNE SÉRIE DES
ANNALES DES MINES
FONDÉES EN 1794

ENJEUX NUMÉRIQUES

Série trimestrielle • N°2 - Juin 2018

Rédaction
Conseil général de l'Économie,
ministère de l'Économie et des Finances
120, rue de Bercy - Télédoc 797
75572 PARIS Cedex 12
Tél. : 01 53 18 52 68
http://www.annales.org

François Valérian
Rédacteur en chef

Gérard Comby
Secrétaire général

Delphine Mantienne
Secrétaire générale adjointe

Liliane Crapanzano
Relectrice

Myriam Michaux
Webmestre

Membres du Comité de Rédaction

Jean-Pierre Dardayrol,
Président du Comité de rédaction
Edmond Baranes
Godefroy Beauvallet
Côme Berbain
Pierre Bonis
Serge Catoire
Michel Cosnard
Arnaud de La Fortelle
Caroline Le Boucher
Alban de Nervaux
Bertrand Pailhès
Grégoire Postel-Vinay
Jacques Serris
Hélène Serveille
Laurent Toutain
Françoise Trassoudaine
François Valérian

Photo de couverture :
Wassily Kandinsky (1864-1944), *Deutliche Verbindung (Liaison manifeste)*.
Aquarelle et encre noire sur papier. Coll. Part.
Photo © CHRISTIE'S IMAGES-BRIDGEMAN IMAGES

Iconographie
Christine de Coninck

Abonnements et ventes
COM & COM
Bâtiment Copernic - 20, avenue Édouard-Herriot
92350 LE PLESSIS-ROBINSON
Alain Bruel
Tél. : 01 40 94 22 22 - Fax : 01 40 94 22 32
a.bruel@cometcom.fr

Mise en page : Nadine Namer
Impression : Printcorp

Éditeur délégué :
FFE – 15, rue des Sablons - 75116 PARIS -
www.ffe.fr

Régie publicitaire : Belvédère Com
Fabrication : Aïda Pereira
aida.pereira@belvederecom.fr
Tél. : 01 53 36 20 46
Directeur de la publicité : Bruno Slama
Tél. : 01 40 09 66 17
bruno.slama@belvederecom.fr

Le sigle « D. R. » en regard de certaines illustrations correspond à des documents ou photographies pour lesquels nos recherches d'ayants droit ou d'héritiers se sont avérées infructueuses.

Big Data : économie et régulation

HORS DOSSIER

Ce numéro a été coordonné par Edmond BARANES

Introduction

Par Edmond BARANES
Professeur d'économie, Université de Montpellier

Après un premier numéro d'*Enjeux numériques* sur l'intelligence artificielle, ce deuxième numéro est consacré au Big Data ; par Big Data, on entend l'accumulation massive de données numériques. C'est un phénomène d'ampleur qui, au-delà des enjeux technologiques, entraîne des changements importants dans l'organisation traditionnelle de l'économie et constitue un défi sociétal.

Le Big Data apparaît comme un élément révélateur de la transformation numérique de nos sociétés. Ce phénomène de prolifération des données modifie notre façon de produire et de consommer, il questionne aussi notre manière de concevoir les libertés individuelles et le principe de souveraineté.

C'est depuis une dizaine d'années, avec l'apparition des équipements connectés (smartphones, tablettes) et des premières applications, que le phénomène de production de données s'est amplifié. Aujourd'hui ce phénomène est renforcé par l'avènement des réseaux sociaux et l'explosion de l'Internet des objets, et se combine avec les progrès en intelligence artificielle.

La « révolution » de la donnée à laquelle nous assistons peut être qualifiée de nouvelle révolution industrielle. La donnée est une ressource importante pour l'économie. Le Big Data permet d'accroître le stock d'informations et le traitement des données permet d'améliorer la qualité de l'information, ce qui accroît ainsi le niveau des connaissances. Ce lien est important car le stock de connaissances joue un rôle fondamental pour l'innovation, la croissance et le développement économique.

Un rapport de l'OCDE publié en 2015 (1) fournit un premier tour d'horizon des effets potentiellement positifs du Big Data sur la croissance économique et le bien-être. Les analyses empiriques sont encore assez limitées et gagneraient certainement à être développées ; l'usage intensif de données permettrait d'expliquer en moyenne annuelle 0,02 % de la croissance entre 2005 et 2012 au Royaume-Uni (2). Concernant l'impact du Big Data sur les performances des entreprises en matière d'innovation, les études se limitent à des analyses sectorielles spécifiques (3) (4). Les résultats confirment toutefois une corrélation significative entre l'usage intensif des données et la productivité des entreprises américaines sur la période 2005-2010, et soulignent une complémentarité entre recours au Big Data et emploi hautement qualifié. Une étude récente (5) appliquée au secteur manufacturier et des services en Allemagne montre que l'usage du Big Data est associé à une plus grande propension des entreprises à innover et que l'usage intensif des données améliore les performances de marché des entreprises innovantes.

Le 28 mars dernier, Cédric Villani remettait au gouvernement son rapport *Donner un sens à l'Intelligence artificielle : pour une stratégie nationale et européenne*. Ce rapport place la donnée au centre de la stratégie qui doit reposer sur « une politique offensive visant à favoriser l'accès aux données,

(1) OCDE (2015), *Data-Driven Innovation: Big Data for Growth and Well-Being*.
(2) GOODRIDGE P. et HASKEL J. (2015), "How Does Big Data Affect GDP? Theory and Evidence for the UK", *Discussion Paper 2015/06*, Imperial College Business School.
(3) BRYNJOLFSSON E. et MCELHERAN K. (2016), "Data in Action: Data-Driven Decision Making in U.S. Manufacturing", *working paper 16-06*, Center for Economic Studies, U.S. Census Bureau.
(4) TAMBE P. (2014), « Big data Investment, Skills, and Firm Value", *Management Science 60(6)*.
(5) NIEBEL T., RASEL F. et VIETE S. (2017), "BIG Data – BIG Gains? Understanding the Link Between Big Data Analytics and Innovation", *Center for European Economic Research*, Mannheim.

la circulation de celles-ci et leur partage ». Il rappelle le caractère non rival de la donnée qui traduit le fait que sa détention et son utilisation par une personne n'empêchent pas d'autres d'en disposer. C'est en ce sens qu'on peut admettre que les données sont pour partie un bien collectif, autrement dit une ressource dont l'usage demande à être défini par la collectivité.

Mais le caractère non rival n'implique pas forcément un accès libre et sans coût aux données. Sous cet angle, les données peuvent apparaître comme une ressource essentielle, ou une infrastructure, ce qui justifie une politique d'ouverture des données afin de libérer les initiatives et de faciliter les innovations. Les données confèrent alors un avantage concurrentiel à ceux qui les détiennent. Aujourd'hui, la place occupée par les géants du numérique que sont les GAFAM (Google, Amazon, Facebook, Apple, Microsoft) leur permet de collecter, de détenir et de valoriser les données. Cela renforce leur position de marché et fragilise les petits acteurs du fait de la relation quasi exponentielle entre la quantité des données collectées et leur valorisation. Dans ce contexte, l'enjeu est de taille pour la France et l'Europe qui cherchent les conditions pour faire émerger leurs champions du numérique.

Pour gagner l'adhésion de la collectivité, une politique de protection des données, en particulier des données personnelles, doit accompagner ce mouvement d'ouverture des données. Le récent Règlement européen sur la Protection des Données (RGPD) permet d'encadrer les conditions de collecte et d'utilisation des données personnelles et doit contribuer à la construction de l'écosystème numérique à l'échelle européenne.

À travers ce numéro d'*Enjeux numériques*, nous souhaitons accompagner la réflexion sur les débats actuels autour du Big Data. La première partie du numéro est articulée autour des enjeux du Big Data. Les contributions sont nombreuses, elles offrent une présentation large des enjeux du Big Data en explorant les aspects économiques, puis réglementaires, et en accordant une place importante aux enjeux de société. La deuxième partie sélectionne des cas d'application empruntés à différents domaines et secteurs de l'économie afin d'illustrer les changements induits par l'essor du Big Data. Ainsi, des exemples sont présentés concernant l'industrie, l'assurance, le marketing prédictif et l'agriculture. Le développement de ces masses de données pose aussi la question de l'articulation et de la complémentarité avec les modes de recueils et de production des données traditionnellement mis en œuvre par les instituts de statistiques : quelles articulations entre Big Data et statistique publique ?

Nous avons enfin souhaité clôturer ce numéro d'*Enjeux numériques* par une interview d'Yves Gassot, pendant plus de vingt ans directeur général de l'IDATE DigiWorld, qui nous offre une vision particulièrement éclairante sur les ruptures introduites par le Big Data.

Big Data : enjeux technologiques et impact scientifique

Par Stephan CLÉMENÇON
Professeur de mathématiques appliquées à Télécom-ParisTech,
Institut Mines-Télécom

L'évocation du terme Big Data provoque généralement une réaction ambivalente. Une crainte, fondée le plus souvent sur des dangers bien réels : une automatisation des processus de décision pouvant s'accompagner d'une perte de contrôle, un impact négatif sur l'emploi, la dépendance de certaines activités à l'égard des systèmes d'information et la disparition de la vie privée. Mais également un engouement certain pour ce que les masses de données aujourd'hui disponibles, combinées à des sciences et technologies de l'information en plein essor, le *machine learning* en particulier, pourraient permettre d'accomplir dans de nombreux secteurs (science, médecine, commerce, transports, communication, sécurité), à l'instar des progrès réalisés ces vingt dernières années dans des domaines tels que la vision par ordinateur ou la reconnaissance de la parole. S'il est encore aujourd'hui difficile de percevoir précisément comment organiser une régulation efficace sans pour autant brider les avancées promises, la maîtrise des risques passe en partie par l'éducation et la formation, par une plus grande diffusion d'une « culture des données et des algorithmes ». Les peurs suscitées par l'automatisation ne sont pas nouvelles. Dans le cas du traitement des masses d'informations numérisées, cette automatisation est pourtant inévitable et souhaitable. Perçue à tort comme une discipline visant à remplacer l'expertise d'un opérateur humain par des machines réalisant des tâches automatisées définies par des données, le *machine learning* a au contraire pour objectif de nous aider à exploiter les données brutes collectées par les capteurs modernes (télescope spatial, spectromètre de masse, téléphones mobiles), portant une information complexe qu'il nous est absolument impossible d'embrasser sans un traitement mathématique adéquat, mis en œuvre au moyen de programmes informatiques dédiés. Il est aujourd'hui à l'œuvre dans de nombreux domaines et s'incarne avec succès dans des applications telles que la vidéosurveillance, la maintenance prédictive des grands systèmes et infrastructures ou les moteurs de recommandation sur le web.

On peut prévoir que ce corpus de connaissances et techniques à l'interface des mathématiques et de l'informatique, en progrès constant depuis quelques décennies, sera encore à l'origine de nombreuses innovations à fort impact sociétal, économique ou scientifique, pour peu que son potentiel soit compris par un public de plus en plus large, qu'il soit maîtrisé par un nombre croissant d'ingénieurs et de cadres techniques, et qu'il se confronte aux enjeux de la société moderne. Le véritable danger de l'automatisation du traitement des données massives résiderait au contraire dans une pénurie d'expertise et des compétences qui permettent de vérifier les conditions dans lesquelles les données sont collectées, d'assurer leur véracité et le bien-fondé des modèles statistiques sur lesquels reposent les applications modernes, et d'interpréter les résultats.

Le paradigme de l'apprentissage statistique

L'un des exemples les plus éloquents de l'impact du Big Data est sans aucun doute celui de la reconnaissance de formes. Celle-ci s'incarne dans les applications de l'intelligence artificielle les plus fréquemment mises en avant aujourd'hui pour illustrer l'efficacité des solutions qu'elle permet de produire, telles que la vision par ordinateur, la reconnaissance automatique de la parole ou de l'écriture manuscrite.

Les concepts mathématiques et algorithmiques à l'œuvre pour mettre au point ces systèmes intelligents sont pourtant loin d'être nouveaux. Même s'ils ont fait l'objet d'une amélioration significative ces dernières décennies, leur élaboration remonte pour l'essentiel à plus d'un demi-siècle. Dans tous ces problèmes, la tâche que la machine doit accomplir consiste, à partir d'une donnée d'entrée X, en la reconnaissance automatique avec une marge d'erreur minimale d'une catégorie Y d'un certain type, spécifié à l'avance, et dont relève la donnée X. Pour reprendre l'exemple de la biométrie, X peut être, par exemple, une image pixellisée ou un signal sonore, et Y est l'identité de l'individu figurant sur l'image ou dont la voix a été capturée par le signal enregistré. Les mêmes technologies sont déployées désormais dans le cadre de l'aide au diagnostic ou pronostic médical ou dans la gestion du risque de crédit, mais on comprendra aisément que les données d'entrée X déterminant alors la catégorie Y dans une bien moindre mesure, le niveau d'erreur attendu est largement supérieur pour ces applications à celui des moteurs de reconnaissance biométrique évoqué précédemment. La reconnaissance de forme est un problème prédictif dans la mesure où les règles élaborées ne doivent pas seulement pouvoir être mises en œuvre au moyen des bibliothèques logicielles disponibles et minimiser l'erreur commise sur une base de données historiques contenant un certain nombre d'exemples (X,Y) appelés « données d'apprentissage », mais ces règles doivent aussi permettre de prédire efficacement le label Y pour de nouvelles entrées X, non encore observées mais issues de la même population statistique que les exemples d'apprentissage (on conviendra qu'il est toujours aisé de « prédire le passé »). On parle alors de capacité de généralisation de la règle prédictive. La formulation du problème d'apprentissage d'une telle règle convoque donc naturellement le langage des probabilités et sa résolution pratique consiste à sélectionner une règle prédictive, au moyen d'un algorithme d'optimisation opérant sur une classe donnée de règles candidates, minimisant une version statistique de la probabilité d'erreur calculée à partir des exemples stockés dans la base d'apprentissage. La théorie mathématique élaborée par Vladimir Vapnik à la fin des années 1960 garantit la capacité de généralisation des règles ainsi construites, pour peu que les classes à partir desquelles l'apprentissage automatique est réalisé soient d'une complexité contrôlée. Le cadre de validité qu'elle a permis de donner à l'apprentissage statistique a fait naître un courant de recherche très actif, mobilisant des chercheurs à l'interface de plusieurs disciplines, les mathématiques et l'informatique bien sûr, mais aussi les sciences cognitives.

L'impact du Big Data

Mais si les concepts fondamentaux du *machine learning* et certains algorithmes tels que les réseaux de neurones sont présents sous des formes très abouties dès la fin des années 1970, ce n'est qu'au commencement de l'ère du Big Data, il y a une dizaine d'années, que le *machine learning* a pu commencer à rencontrer le succès qu'on lui connaît aujourd'hui. Les obstacles principaux résidaient d'une part en la rareté de l'information numérisée, la collection de données s'effectuant alors le plus souvent à travers des plans de sondage coûteux, et d'autre part en des capacités de mémoire et de calcul limitées, interdisant la mise en œuvre de programmes d'optimisation opérant sur de vastes classes de règles pour réaliser un apprentissage efficace. Dans bien des situations, les faibles capacités prédictives des règles produites par le *machine learning* pouvaient ainsi être imputées tout à la fois à une erreur statistique inhérente au faible nombre d'exemples à partir desquels l'apprentissage s'effectue, et au caractère fruste des modèles prédictifs constituant les classes sur lesquelles les programmes d'optimisation peuvent être appliqués. Les briques technologiques ayant permis le développement du web, comme les systèmes de fichiers distribués du *framework* Hadoop ou les langages de programmation tels que MapReduce, ont en effet engendré des progrès considérables dans le domaine de la collecte et du stockage de données et du traitement massivement distribué et parallélisé. Les mégadonnées du web, les immenses bibliothèques d'images, de sons ou de textes « étiquetés » auxquelles il permet d'accéder, entraînent ainsi les moteurs de reconnaissance de contenu avec d'innombrables exemples. Les avancées réalisées dans la gestion

de la mémoire ou dans le domaine du calcul parallélisé grâce en particulier aux processus graphiques permettent la mise en œuvre de programmes d'apprentissage opérant sur des classes très flexibles, telles que les réseaux de neurones profonds (*deep learning*), susceptibles, pour de nombreux problèmes, de rendre compte très efficacement de la façon dont l'information en entrée X permet de prédire la sortie Y. L'ubiquité des capteurs et le développement de l'*Internet of Things* (IoT) facilitent désormais l'accès à l'information numérique, et d'innombrables applications sont développées aujourd'hui sur le modèle de la reconnaissance de forme.

Les infrastructures de collecte, de gestion des masses de données et de calcul ne conditionnent cependant pas à elles seules les progrès réalisés dans le domaine du *machine learning*, et l'avenir ne se bornera pas à simplement décliner les applications du *deep learning*. Par exemple, la volonté d'embarquer des moteurs de reconnaissance biométrique performants dans des smartphones sans compromettre leur autonomie incite les chercheurs à comprendre comment « compresser » ces réseaux profonds de manière à limiter les échanges d'énergie sans pour autant dégrader la qualité de la reconnaissance.

Si le Big Data correspond pour l'apprentissage statistique à une sorte de nirvana, dont les méthodes sont d'autant plus fiables qu'elles sont fondées sur l'observation d'expériences « en grand nombre », le contrôle des conditions d'acquisition des données et des hypothèses de validité des algorithmes prédictifs est indispensable au succès des modèles calculés par les machines. La culture probabiliste et statistique devrait ainsi prendre une place de plus en plus importante dans la plupart des cursus universitaires, et pas seulement dans celui des *data scientists*, ces nouveaux spécialistes des statistiques algorithmiques. La diffusion accrue de cette culture ferait en particulier s'évanouir la crainte d'un monde où le Big Data permettrait de prédire sans erreur nos comportements ou la date de notre mort... Les « grands nombres » permettent en effet d'estimer la performance prédictive des modèles, d'évaluer les risques avec précision et d'optimiser les décisions en univers incertain, mais pas de réduire le caractère intrinsèquement aléatoire de certains phénomènes.

Modèles économiques des données : une relation complexe entre demande et offre

Par Paul BELLEFLAMME
Professeur à Aix-Marseille School of Economics

Le 19 mars 2018, Facebook dégringole en Bourse après la révélation que la société Cambridge Analytica a utilisé les données personnelles de près de 50 millions d'utilisateurs du réseau social sans leur consentement [(1)]. On apprend aussi ce jour-là que les *selfies* des internautes, qui peuvent servir à valider des processus d'identification, se négocient à des prix allant jusqu'à 70 dollars sur les marchés clandestins du « Dark Web ». Quelques mois plus tôt, la société iRobot revenait sur ses déclarations antérieures selon lesquelles elle cherchait à revendre les données collectées par ses robots aspirateurs Roomba, capables de réaliser une carte virtuelle des endroits qu'ils nettoient. La société Uber, quant à elle, annonçait la création d'une plateforme visant à partager gratuitement des données de déplacements de ses chauffeurs et clients avec les planificateurs urbains des quatre cent cinquante villes où elle est active [(2)].

Ces quelques événements récents montrent toute l'importance que les échanges de données occupent dans nos économies. Ils illustrent aussi les différentes formes que peuvent prendre ces échanges, partage librement consenti (cas d'Uber), vol pur et simple (les *selfies* sur le Dark Web), ou encore échanges encadrés par des dispositions contractuelles plus ou moins claires (cas de Facebook et d'iRobot).

L'objectif de cet article est de mieux faire comprendre comment s'organisent les échanges de données. Pour ce faire, nous commençons par décrire le côté de la demande, en étudiant pourquoi, et comment, les données acquièrent de la valeur (Section 1). Nous considérons ensuite le côté de l'offre, en nous demandant d'où viennent les données et qui en contrôle la production et la collecte (Section 2). Il s'agit enfin de comprendre comment l'offre et la demande se rencontrent (Section 3). Nous concluons en réfléchissant aux évolutions que pourraient prendre les échanges de données dans le futur.

Le côté de la demande

La demande de données émane d'entreprises, mais aussi d'organisations non commerciales (des villes par exemple) qui cherchent à améliorer leurs pratiques. Il s'agit d'une demande induite car ce ne sont pas les données en elles-mêmes qui sont recherchées mais bien les informations qui peuvent en être extraites et, finalement, les connaissances que génèrent ces informations et qui contribuent à la prise de décisions [(3)]. On comprend donc pourquoi la demande de données est un phénomène récent. En effet, la capacité d'accroître la valeur des données en les transformant en informations a considérablement augmenté ces dernières années sous l'effet conjoint de la numérisation et de la « datafication ». La première tendance est la généralisation du format numérique qui permet de stocker, dupliquer et transmettre les données électroniquement bien plus vite et à un coût énergétique nettement moindre. La seconde tendance est la multiplication des traces

(1) Révélations faites par *The Guardian* http://bit.ly/2plU1sM) et *The New York Times* (http://nyti.ms/2u1nLjw
(2) Voir, respectivement, http://bit.ly/2u8ys3W, http://bit.ly/2DFoxCv et http://bit.ly/2prPK6B
(3) Thierauf (1999) définit les données comme une collection non structurée de faits et de chiffres, l'information comme des données structurées et la connaissance comme de « l'information à propos de l'information ».

numériques laissées derrière elles par nos activités, que ce soit par nos ordinateurs et smartphones, par les réseaux sociaux, ou les senseurs de nos objets connectés. À cela s'ajoute le développement d'une nouvelle discipline scientifique, la science des données, qui combine outils mathématiques, statistiques et informatiques pour optimiser l'extraction de connaissances à partir d'ensembles de données.

En résumé, la demande de données est en pleine expansion parce que tant les données disponibles que la capacité de les traiter ne cessent de croître. La valeur des données augmente en effet avec ce qu'il est convenu d'appeler les quatre V des données, à savoir leur volume (d'où le terme de *big data*, qui suggère des économies d'échelle), leur variété (c'est-à-dire la diversité de leurs sources, qui suggère des économies d'envergure), leur vélocité (c'est-à-dire la rapidité avec laquelle les flux de données peuvent être traités) et, naturellement, leur véracité (ou leur précision, qui détermine la confiance qu'on peut leur accorder).

Les entreprises sont avides de données parce qu'elles cherchent à améliorer leurs processus de production, à développer des produits et services innovants et à mieux cibler leurs clients avec des offres, des publicités et des prix adaptés. Comme chaque entreprise a pour objectif de surpasser ses concurrentes, une course s'engage à qui utilisera au mieux les données disponibles. En découlent deux conséquences importantes pour la demande de données. D'une part, les entreprises ont une disposition à payer beaucoup plus élevée pour des données auxquelles elles ont un accès exclusif que pour des données qu'elles devraient partager avec leurs concurrents. D'autre part, il est possible qu'au sein d'une industrie, les entreprises concurrentes investissent de manière excessive dans l'acquisition et le traitement de données, avec comme effet que les profits des entreprises finissent par baisser. En d'autres termes, comme dans le célèbre dilemme du prisonnier, des entreprises concurrentes gagneraient à restreindre collectivement leur utilisation de données, mais aucune n'y trouve intérêt individuellement.

Le côté de l'offre

Les données que valorisent les entreprises proviennent de trois sources. Tout d'abord, de nombreuses bases de données sont en accès libre. La plus grosse partie de ces « données ouvertes » est produite par le secteur public (on pense à des données statistiques, scientifiques ou cartographiques) ; des organisations comme les universités ou les organisations non gouvernementales ouvrent également leurs données ; même des entreprises commerciales peuvent y trouver un intérêt (à l'instar d'Uber mentionnée dans l'introduction). Ensuite, les entreprises produisent elles-mêmes énormément de données au fil de leurs activités et par les produits qu'elles vendent [4]. Finalement, vous et moi sommes sans doute les plus gros pourvoyeurs des données qui intéressent les entreprises. C'est le phénomène de « datafication » que nous évoquions plus haut : nous produisons des données, soit directement par nos activités (les photos ou les « likes » que nous postons sur les réseaux sociaux, les sites que nous visitons, les mails que nous envoyons, etc.), soit indirectement par les machines ou équipements que nous utilisons (un smartphone dont la géolocalisation est activée ou une montre connectée par exemple). Ces données sont précieuses pour les entreprises dans la mesure où elles indiquent nos goûts, nos habitudes de consommation, nos interactions sociales, etc.

Pour la suite de notre analyse, il est important de déterminer dans quelle mesure on peut parler d'une offre de données. Pour qu'une offre existe, il faut que l'accès aux données puisse être contrôlé, de sorte que le producteur puisse fixer les termes d'une éventuelle transaction. Pour les deux

(4) On estime par exemple qu'une voiture autonome génère jusqu'à 100 gigabytes de données par seconde (soit l'équivalent de plus de 5 millions de pages de texte).

premières sources de données, les producteurs – organismes publics, entreprises – sont largement en mesure de déterminer les conditions d'accès à leurs données : l'accès est délibérément rendu public pour les données ouvertes ; pour les données d'entreprises, nous verrons dans la section suivante que l'accès est le plus souvent fermé ou encadré par des dispositions contractuelles.

Qu'en est-il des données que nous produisons en tant qu'individus ? Pouvons-nous en contrôler l'accès ? En théorie, oui. Les sites web que nous visitons, ou les objets connectés que nous utilisons, nous invitent à signifier notre accord avec leurs conditions d'utilisation. Même si la possibilité nous est laissée de refuser que nos données soient collectées, nous n'exerçons pas, ou très peu, cette option (qu'on appelle en anglais « opt out »). Pourquoi ? Une première raison est que nous jugeons trop coûteux (en temps et en effort) de prendre connaissance des conditions d'utilisation ou d'appliquer des mesures pour limiter la collecte de nos données [(5)]. Une seconde raison, qui justifie partiellement la première, est que nous acceptons d'obtenir, en échange de nos données, des services moins chers (souvent gratuits), mieux adaptés à nos besoins (comme des offres ciblées) et potentiellement de meilleure qualité [(6)]. Cela revient à dire que nous associons un « prix virtuel » à nos données et donc à notre vie privée.

Il arrive que ce prix virtuel devienne un prix réel. C'est le cas quand des entreprises (par exemple des fournisseurs d'accès à Internet) différencient leurs services en proposant aux consommateurs de payer plus cher pour éviter de voir leurs données collectées ou de recevoir des publicités ciblées. En choisissant ce genre d'offres, les consommateurs révèlent leur volonté de payer pour protéger leur vie privée. Il s'agit toujours ici d'un système « opt out », puisque c'est au consommateur de payer pour fermer l'accès à ces données. Que se passe-t-il si, à l'inverse, c'est à l'entreprise de payer le consommateur pour qu'il ouvre l'accès à ses données (système « opt in ») ? On a envie de penser que rien ne devrait changer pour un même montant monétaire (à payer ou à recevoir) et une même variation (à la hausse ou à la baisse) du degré de protection des données. Mais des études montrent qu'en général, les consommateurs demandent en échange d'une érosion de leur vie privée un montant monétaire plus élevé que celui qu'ils sont prêts à payer pour protéger leur vie privée dans une même mesure [(7)]. Les consommateurs semblent donc attacher une valeur plus importante à leurs données quand leur consentement est nécessaire pour l'utilisation (« opt in »), plutôt que pour l'absence d'utilisation (« opt out ») de celles-ci.

La mise en relation de l'offre et de la demande

Williamson (1991) distingue trois façons d'organiser les transactions économiques : la « hiérarchie » organise les transactions au sein d'une entreprise intégrée, le « marché » utilise le mécanisme des prix pour coordonner offre et demande et, entre ces deux extrêmes, les « formes hybrides » reposent sur des contrats spécifiques.

Actuellement, les transactions sur les données s'organisent essentiellement par le mode hiérarchique ou par des formes hybrides. Dans le premier cas, les entreprises collectent directement, ou produisent elles-mêmes, les données dont elles ont besoin. Quand il s'agit de données personnelles, nous avons vu plus haut que la collecte s'appuie sur des contrats de type « opt out » : pour dire les choses crûment, les entreprises se servent tant que les consommateurs ne les en empêchent pas. Nous avons montré aussi que l'intégration verticale se justifie dès lors que les données permettent

(5) Il faudrait 76 jours pour lire l'intégralité des conditions d'utilisation qu'un Américain moyen accepte de signer en un an (GRALLET *et al.*, 2018). Pour limiter l'accès à ses données, il est possible, par exemple, d'effacer les cookies de son navigateur ou de passer par des serveurs proxy.
(6) Par exemple, un compteur communicant n'a véritablement de valeur ajoutée que s'il peut mesurer de manière précise notre consommation d'eau ou d'électricité.
(7) Voir ACQUISTI *et al.* (2013) ; voir aussi SCHOLZ (2014).

d'obtenir un avantage concurrentiel (les entreprises n'ont en effet aucun intérêt à partager les données qu'elles récoltent et moins encore l'information et la connaissance qu'elles en extraient). Les dispositions légales limitant le partage de données personnelles viennent renforcer cette tendance à l'intégration verticale.

Il arrive toutefois que des entreprises trouvent profitable de partager leurs données, afin de mieux coordonner leurs activités [8]. Les transactions se basent alors sur des contrats multilatéraux de long terme. Une autre forme de gouvernance hybride est le recours à des intermédiaires spécialisés dans la collecte et le traitement de données. On les appelle « courtiers en données » (*data brokers* en anglais). Les services sur mesure que proposent ces courtiers sont particulièrement prisés par les entreprises qui ne peuvent pas collecter de données par elles-mêmes. En raison des économies d'échelle et d'envergure évoquées plus haut, l'industrie des courtiers en données est dominée par quelques entreprises, en majorité américaines, qui rassemblent des données diverses sur des centaines de millions de consommateurs de par le monde ; citons Acxiom (marketing), Equifax (assurance), Experian (crédit), Corelogic (immobilier) ou Datalogix (finance).

À ce jour, il n'existe pas de « marché des données » à proprement parler. On trouve certes quelques plateformes d'échange de données mais celles-ci sont limitées à une industrie particulière et restreignent considérablement les transactions qui peuvent être effectuées. Ceci s'explique par le paradoxe suivant : comme les données sont stratégiques, la disposition à payer pour des données non exclusives est généralement faible, voire nulle ; mais parce que les données sont non rivales (la consommation par l'un ne réduit pas les possibilités de consommation par l'autre), l'exclusivité est difficile à garantir, singulièrement dans un mécanisme d'échange décentralisé. En outre, il est difficile d'établir rigoureusement la véracité des données, ainsi que leur valeur, en raison de leur unicité (absence de point de comparaison) ou de leur complémentarité (il faut combiner plusieurs bases de données pour extraire de l'information pertinente [9]).

Conclusion

En résumé, une quantité sans cesse croissante de données est produite, collectée et utilisée mais, en définitive, une fraction assez limitée de ces données est échangée. Nous avons identifié trois explications : le caractère stratégique des données pour les entreprises, la difficulté d'organiser des places de marché décentralisées et le manque de contrôle des individus sur les données qu'ils produisent. Sur ce dernier point, on peut s'attendre à des transformations importantes dans un futur assez proche. En effet, un nouveau texte européen, intitulé Règlement général sur la Protection des Données (RGPD), vient d'entrer en vigueur ; il impose aux entreprises de donner aux individus davantage de contrôle sur leurs données personnelles. Cela signifie que les entreprises vont devoir obtenir un consentement explicite et positif des individus pour pouvoir utiliser leurs données et, également, assurer la portabilité de ces données (c'est-à-dire permettre aux consommateurs d'emmener leurs données avec eux lorsqu'ils changent de fournisseur). Comme l'expliquent Peitz et Schweitzer (2017), la portabilité empêche le verrouillage et facilite ainsi la concurrence dans l'accès aux données personnelles (à défaut de mettre en place un véritable marché secondaire des données).

Une autre source de changement est le développement de nouveaux intermédiaires qui proposent aux consommateurs des solutions pour gérer activement leurs données personnelles et, potentiellement, les monétiser [10]. Enfin, le scandale Facebook/Cambridge Analytica (qui ouvre cet article)

(8) On pense aux entreprises collaborant à la mise au point de voitures autonomes.
(9) Pour une analyse plus détaillée, voir KOUTROUMPIS *et al.* (2017).
(10) On les appelle PIMS (*Personal Information Management Systems*) ou systèmes de gestion des informations personnelles. Les plus connus sont Datacoup, Digi.me et Meeco.

a suscité de telles réactions des internautes et des pouvoirs publics qu'on est en droit de penser qu'une nouvelle ère commence, où les transactions sur les données personnelles seront plus encadrées, plus transparentes et plus respectueuses des individus.

Références

ACQUISTI A., JOHN L.K. & LOEWENSTEIN G. (2013), "What is Privacy Worth?", *The Journal of Legal Studies*, 42, pp. 249-274.

GRALLET G., PONCET G. & PONS H. (2018), « Données personnelles : comment reprendre le contrôle ? », *Le Point* (25 janvier).

KOUTROUMPIS P., LEIPONEN A. & THOMAS L. (2017), "The (Unfulfilled) Potential of Data Marketplaces", *ETLA Working Papers*, n° 53. http://pub.etla.fi/ETLA-Working-Papers-53.pdf

PEITZ M. & SCHWEITZER H. (2017), "Datenmärkte in der digitalisierten Wirtschaft: Funktionsdefizite und Regelungsbedarf", Discussion Paper No. 17-043, ZEW, Mannheim.

SCHOLZ E.-M. (2014), "Putting a Price on Privacy – an Introduction to the Economics of Privacy", IPdigIT.eu (1 June), http://www.ipdigit.eu/2014/06/putting-a-price-on-privacy-an-introduction-to-the-economics-of-privacy/

THIERAUF R.J. (1999), *Knowledge Management Systems*, Quorum Books.

WILLIAMSON O.E. (1991), "Comparative Economic Organization: the Analysis of Discrete Structural Alternatives", *Administrative Science Quarterly*, 36, 269-296.

Vie privée, valeur des données personnelles et régulation

Par Grazia CECERE
IMT, TEM
et Matthieu MANANT
Université Paris Sud, RITM

Le rôle des données personnelles en économie a été accentué par la numérisation croissante de l'économie, car il est désormais possible de collecter, stocker et traiter des quantités énormes de données à des coûts de plus en plus faibles. Si l'exploitation des données personnelles permet aux entreprises de faire aux utilisateurs de meilleures offres en aidant à réduire les coûts de recherche, elle peut également être à l'origine de discriminations préjudiciables ou de sollicitations non souhaitées et inappropriées pour ces mêmes utilisateurs (Acquisti *et al.*, 2017). Cette utilisation secondaire des données personnelles peut survenir sans que l'utilisateur n'en ait conscience. D'un point de vue économique, cette exploitation possible des données personnelles est liée au fait qu'il s'agit d'un bien d'information non rival et non excluable (Cecere *et al.*, 2017). La littérature académique en économie de la vie privée s'intéresse au lien entre les comportements de divulgation de données personnelles des individus et les stratégies d'innovation des entreprises lorsqu'elles utilisent ces données. L'omniprésence de l'Internet – y compris celle de l'Internet des objets – a accru l'importance des données dans tous les secteurs, et plus particulièrement dans la publicité, ainsi que sur les sites d'e-commerce et les plateformes en ligne. Récemment, une littérature académique importante en économie et en marketing a commencé à s'intéresser aux comportements individuels dans différents contextes (Acquisti *et al.*, 2012) et, en particulier, à étudier l'efficacité de la publicité personnalisée des entreprises (Lambrecht et Tucker, 2013) et l'impact de la réglementation sur la protection de la vie privée (Campbell *et al.*, 2015).

Quelle valeur des données personnelles ?

Les entreprises de l'Internet parviennent de mieux en mieux à cibler leurs utilisateurs par une exploitation de plus en plus fine de leurs données personnelles. Si la valeur des données personnelles ainsi générée par les entreprises augmente sans cesse, la mesure ou l'estimation de cette valeur restent en pratique encore difficiles. L'analyse économique des choix et des arbitrages des individus et des entreprises sur les questions de vie privée donne une première appréciation de cette valeur. Pour les individus, le choix de partager leurs données sur les plateformes en ligne permet d'accéder à des services personnalisés et souvent gratuits (Acquisti, 2010). La personnalisation de ces services est souvent améliorée par la fréquentation de ces plateformes, les utilisateurs bénéficiant alors d'externalités liées à la divulgation de données par les autres utilisateurs. Par exemple, en faisant des recommandations – bonnes ou mauvaises – pour des produits qu'ils connaissent, les utilisateurs de sites Internet révèlent leurs préférences et ces informations peuvent ensuite bénéficier aux autres utilisateurs qui ont alors accès à cette information, soit directement, soit sous forme agrégée. La mise à profit de la divulgation de données personnelles n'est cependant pas le seul moyen pour les entreprises d'en recueillir. Les entreprises peuvent en effet s'appuyer sur les données d'activités, comme la navigation ou la localisation, pour prévoir les caractéristiques de leurs clients (Kosinski *et al.*, 2013). Les données personnelles collectées sont alors très diverses et directement personnelles, comme l'âge, l'adresse, le sexe, des préférences révélées par les médias sociaux, les achats ou des commentaires, etc. Ces données peuvent

engendrer encore plus de valeur lorsqu'elles sont combinées avec de grandes bases d'autres données (Lambrecht et Tucker, 2017 ; *The Economist,* 2010).

Les littératures académiques en organisation industrielle et en marketing étudient principalement la façon dont les entreprises exploitent les données personnelles et la manière dont cette exploitation peut créer ou stimuler de nouveaux modèles d'affaires, et donc générer de l'innovation. Les collectes de bases de données très volumineuses, l'analyse de ces données et les progrès du marketing permettent une exploitation de ces données à un niveau sans précédent et qui s'améliore régulièrement (Acquisti, 2014). La valeur des données repose donc sur la façon dont les entreprises les intègrent à leur business model. La littérature théorique repose largement sur l'hypothèse que les agents – individus et entreprises – attribuent une valeur aux données, et que les stratégies des entreprises sont affectées par les quantités et la qualité des données. Ainsi, les modèles théoriques soulignent à quel point les stratégies et modèles d'affaires des entreprises reposent sur la divulgation de données personnelles par les utilisateurs, et donc sur les préférences des consommateurs dans leur vie privée. Plus généralement, ces approches théoriques supposent une acceptation par les utilisateurs du modèle de l'Internet – services gratuits en échange de données personnelles. Fudenberg et Tirole (1998) montrent ainsi qu'un monopole vendant des biens durables peut adopter des stratégies différentes selon les types de consommateurs – anonyme, semi-anonyme ou identifié. Cependant, à notre connaissance, peu de travaux empiriques cherchent à estimer la valeur réelle des données personnelles pour les entreprises, bien que ces données soient clairement valorisées par les entreprises. Un rapport de l'OCDE (2013) suggère toutefois différentes méthodes détaillées d'évaluation des données personnelles et propose un aperçu des entreprises qui opèrent sur le marché en incluant les sociétés de l'Internet directement en lien avec les utilisateurs, mais également des entreprises tierces qui échangent et traitent ces données. En particulier, le rapport s'intéresse à la valeur des données sur les marchés des *data brokers* en montrant que la valeur des données personnelles est liée à la qualité des données collectées comme à leur volume.

Comment les firmes valorisent-elles les données ?

Pour Acquisti *et al.* (2016), trois marchés liés aux données personnelles coexistent : (1) un marché où les entreprises proposent des services aux utilisateurs en échange de leurs données personnelles, (2) un marché où les individus paient pour se protéger, et (3) un marché des *data brokers* – des sociétés tierces – où des entreprises collectent des données et les utilisent pour une commercialisation en *Business-to-Business* (B-to-B).

Sur le marché primaire qui regroupe les sociétés de l'Internet et leurs clients, les données personnelles permettent de concevoir des campagnes publicitaires plus efficaces et de fixer des prix proches de la disposition à payer des individus. Pour la plupart des entreprises en ligne, les revenus publicitaires représentent une source majeure de revenus (Martin et Murphy, 2016). L'exploitation des données de navigation, qui fournissent des informations détaillées sur la façon dont les individus interagissent avec les sites web et la publicité, ainsi que l'utilisation croissante des algorithmes contribuent dans un second temps à accélérer le rythme de l'innovation. Varian (1997) souligne ce double usage des données personnelles par les sociétés de l'Internet : la première utilisation facilite les interactions des entreprises avec leurs clients, alors que la seconde utilisation peut supposer une transmission de ces données à une ou plusieurs entreprises tierces mieux à même d'exploiter les données personnelles. L'exploitation des données personnelles permet une discrimination par les prix de premier degré ou, de façon plus réaliste, une discrimination de troisième degré [(1)], car

(1) La discrimination du premier degré est permise par la connaissance des préférences de chaque individu, celle du deuxième degré par la connaissance des préférences de sous-groupes non identifiés, et celle du toisième degré par la connaissance des préférences de sous-groupes identifiés.

ces données permettent à une entreprise d'identifier le prix de réserve d'un individu. Dans cette perspective, le lien entre les stratégies des entreprises et les données personnelles, et donc la vie privée, apparaît central (Taylor, 2004). En suivant cette hypothèse d'un effet positif de la discrimination pour les utilisateurs, un travail théorique récent de Belleflamme et Vergote (2016) suggère que l'utilisation de technologies pour dissimuler des informations personnelles pourrait réduire le surplus du consommateur. Outre la discrimination par les prix, d'autres formes de discrimination préjudiciables aux individus peuvent cependant également exister, ce qui pose la question de l'usage fait des données personnelles et donc des conséquences sur la confiance des utilisateurs. Ainsi, des travaux récents montrent que, sur le marché du travail, les recruteurs peuvent discriminer entre candidats sur la base des informations trouvées sur les médias sociaux (Acquisti et Fong, 2015 ; Manant *et al.*, 2015). Lambrecht et Tucker (2018) ajoutent à ces travaux le rôle des algorithmes utilisés par les médias sociaux qui reproduisent les discriminations hors ligne en s'appuyant sur des bases de données elles-mêmes biaisées. Plus globalement, cette littérature académique souligne comment les entreprises peuvent exploiter les informations personnelles que les utilisateurs « laissent » en ligne.

Sur le marché secondaire qui regroupe les sociétés de l'Internet et les tiers, l'exploitation des données personnelles est réalisée à l'échelle du marché quand les données personnelles sont achetées par des sociétés tierces, telles que de nombreuses sociétés de marketing comme BlueKai et Avarto spécialisées dans la gestion de données. L'utilisation secondaire de données personnelles par des tiers survient lorsqu'elles sont transmises à des sociétés telles que des courtiers de données, des agrégateurs de données, des annonceurs ou, plus largement, à des entreprises qui ont la compétence de les exploiter (Akçura et Srinivasan, 2005). L'utilisation par des tiers et l'utilisation secondaire de données personnelles au sein d'une même entreprise semblent moins légitimes si des données personnelles sont envoyées ou utilisées à l'insu de l'utilisateur.

Quelle valeur des données pour la régulation ?

L'émergence de nouvelles entreprises dont les modèles d'affaires sont basés sur l'exploitation des données personnelles a attiré l'attention des régulateurs sur la nécessité de trouver un juste équilibre entre protection de la vie privée et promotion du partage de données pour encourager l'innovation et améliorer les services (Tucker, 2012). La définition de la source de valeur des données est indispensable pour bien définir l'impact de la régulation sur les stratégies des entreprises. L'intervention publique sur les questions de vie privée reste cependant complexe. Premièrement, l'innovation dans les secteurs où les données personnelles jouent un rôle-clé est difficile et la concurrence entre les acteurs est forte. Deuxièmement, l'élaboration d'une réglementation visant à protéger la vie privée des individus peut avoir une influence sur les choix de ces mêmes individus (Marthews et Tucker, 2017), mais également sur les choix des entreprises (Miller et Tucker, 2009). L'examen des instruments utilisés par les autorités de réglementation permet de comprendre comment la régulation sur les questions de vie privée peut avoir des effets sur les fonctionnements des marchés.

Aux États-Unis, où la Federal Trade Commission (FTC) fournit des directives au niveau sectoriel, c'est l'autorégulation qui prévaut et le principe est de stimuler la « concurrence sur la vie privée » tout en tenant compte des défaillances de marché (Cecere et Rochelandet, 2012). Cette approche considère les deux côtés du marché en supposant que les consommateurs peuvent décider de garder confidentielles leurs préférences en matière de vie privée et que les entreprises se conforment au respect du principe de transparence et de contrôle des questions de vie privée, par exemple en fournissant des informations détaillées sur la collecte des données. Dans cette perspective, les politiques de confidentialité, souvent au travers des chartes de vie privée, reposent sur un principe de notification et de consentement selon lequel les individus sont censés lire les politiques de confidentialité des entreprises et choisir de consentir, ou non, aux conditions de service (Cranor, 2012). Dans ce cadre de régulation, les politiques de confidentialité doivent fournir des informa-

tions suffisantes aux individus sur la manière dont les entreprises collectent, utilisent, partagent et sécurisent les données personnelles (Marotta-Wurgler, 2016). Cependant, des preuves empiriques montrent que ces politiques sont trop longues à lire et trop complexes à comprendre pour un non-spécialiste (McDonald et Cranor, 2008). L'approche de la FTC a ainsi encouragé la création de services de certification en ligne par des tiers privés comme TRUSTe et BBB avec leurs labels ou sceaux de vie privée qui aident à réduire, pour les utilisateurs, les coûts cognitifs liés à l'évaluation des risques. Néanmoins, le recours à ces services soulève un problème de sélection adverse. Des recherches empiriques montrent en effet que les sites Internet certifiés par TRUSTe sont plus de deux fois davantage susceptibles de ne pas être dignes de confiance que les sites non certifiés (Edelman, 2011). Ces résultats suggèrent la nécessité d'une intervention du régulateur pour assurer la qualité des sceaux privés.

En Europe, l'approche réglementaire est davantage axée sur la mise en place d'un cadre général visant à protéger la vie privée des consommateurs dans tous les secteurs, ce qui est une approche notablement différente de l'approche américaine qui s'appuie principalement sur l'autorégulation. Il s'agit dans le cas européen de limiter les effets de sélection adverse en garantissant davantage aux individus un cadre rigoureux de protection des données personnelles que doivent respecter les entreprises. Le Règlement général sur la Protection des Données (RGPD), entré en vigueur le 25 mai 2018, impose désormais aux entreprises l'obligation de demander aux consommateurs leur consentement pour utiliser leurs données, et permet aux individus d'accéder à plus d'informations sur la façon dont leurs données sont traitées par les entreprises. Il renforce également le droit des individus à l'oubli de leurs données, c'est-à-dire que ces derniers peuvent demander la suppression ou la modification de leurs données à ceux qui les détiennent. La directive européenne encourage la protection de la vie privée dès la conception (*privacy-by-design*), en favorisant la prise en compte des questions de vie privée dès les premières phases, puis tout au long du développement d'un bien ou d'un service. Ce règlement, adopté en 2016, a remplacé le précédent cadre législatif européen en mettant à jour la précédente directive 95/46/CE sur la protection des données et la directive sur la vie privée et les communications électroniques 2002/58/CE. Globalement, si l'approche européenne est censée apporter aux individus plus de transparence et de protection qu'aux États-Unis, elle devrait également être plus coûteuse pour les entreprises qui s'y conforment, car elle leur impose des obligations plus fortes.

Bien que la réglementation de la vie privée s'adresse aux consommateurs et aux entreprises, elle peut également avoir des conséquences indirectes sur la structure du marché. Dans l'ensemble, la direction et la taille de ces effets ne sont pas claires. Alors que la réglementation aide à créer un cadre clair pour les entreprises, il apparaît nécessaire de comprendre le rôle global joué par les données personnelles sur les marchés. Pour cette raison, nous nous sommes concentrés ici sur les principes régissant la réglementation de la vie privée et sur les preuves théoriques et empiriques de l'impact de cette réglementation sur les marchés, mais aussi sur la façon dont les technologies et les violations de données personnelles peuvent influencer ces marchés.

Références

ACQUISTI A. (2010), "The Economics of Personal Data and the Economics of Privacy", OECD Conference Centre.

ACQUISTI A. (2014), "From the economics of privacy to the economics of big data", *Privacy, Big Data, and the Public Good: Frameworks for Engagement*; Stefan Bender, Julia Lane, Helen Nissenbaum, and Victoria Stodden (eds), 76-95.

ACQUISTI A. & M. FONG C. (2012), "An Experiment in Hiring Discrimination Via Online Social Networks", *SSRN Electronic Journal.*

ACQUISTI A., JOHN L. K. & LOEWENSTEIN G. (2012), "The Impact of Relative Standards on the Propensity to Disclose", *Journal of Marketing Research*, 49, 160-174.

ACQUISTI A., TAYLOR C. & WAGMAN L. (2016), "The economics of privacy", *Journal of economic Literature*, 54(2), 442-492.

AKÇURA M. T. & SRINIVASAN K (2005), "Research note: customer intimacy and cross-selling strategy", *Management Science*, 51(6), 1007-1012.

BELLEFLAMME P. & VERGOTE W. (2016), "Monopoly price discrimination and privacy: The hidden cost of hiding", *Economic Letters*, 149, 141-144.

CAMPBELL J., GOLDFARB A. & TUCKER C., "Privacy regulation and market structure", *Journal of Economics & Management Strategy* 24 (1), 47-73.

CECERE G., LE GUEL F., MANANT M. & SOULIÉ N. (2017), "The economics of privacy", *The New Palgrave Dictionary of Economics*.

CECERE G. & ROCHELANDET F. (2012), « Modèle d'affaires numériques, données personnelles et sites web. Une évaluation empirique », *Revue française de gestion*, 224, 111-124.

CRANOR L. F. (2012), "Necessary but not sufficient: Standardized mechanisms for privacy notice and choice", *J. on Telecomm. and High Tech. L.,* 10, 273-307.

EDELMAN B. (2011), "Adverse Selection in Online 'Trust' Certifications and Search Results", *Electronic Commerce Research and Applications*, 10(1), 17-25.

FUDENBERG D. & TIROLE J. (1998), "Upgrades, tradeins, and buybacks", *RAND Journal of Economics*, 29(2), 235-258.

KOSINSKI M., STILLWELL D. & GRAEPEL T. (2013), "Private traits and attributes are predictable from digital records of human behavior", *Proceedings of the National Academy of Sciences*, 110(15), 5802-5805.

LAMBRECHT A., GOLDFARB A., BONATTI A., GHOSE A., GOLDSTEIN D., LEWIS R., RAO A., SAHNI N. & YAO S. (2014), "How do firms make money selling digital goods online?", *Marketing Letters*, 25, 331-341.

LAMBRECHT A. & TUCKER C.E. (2013), "When does retargeting work? Information specificity in online advertising", *Journal of Marketing research*, 50(5), 561-576.

LAMBRECHT A. & TUCKER C.E. (2017), "Can Big Data protect a firm from competition?", *CPI Antitrust Chronicle*.

LAMBRECHT A. & TUCKER C.E. (2018), "Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads", *Social Science Research Network. WP.*

MANANT M., PAJAK S. & SOULIÉ N. (2018), "Can Social Media lead to Labour Market Discrimination: A Field Experiment", *WP*.

MAROTTA-WURGLER F. (2016), "Self-regulation and competition in Privacy Policies", *Journal of Legal Studies*, 45(S2), 13-39.

MARTHEWS A. & TUCKER C. (2017), "Government Surveillance and Internet Search Behavior", *SSRN*.

MARTIN K. D. & MURPHY P. E. (2016), "The role of data privacy in marketing", *Journal of the Academy of Marketing Science*, forthcoming.

McDONALD A. M. & CRANOR L. F. (2008), "The Cost of Reading Privacy Policies", I/S: A *Journal of Law and Policy for the Information Society*, 4(3), 540-565.

MILLER A. R. & TUCKER C. (2009), "Privacy protection and technology diffusion: The case of electronic medical records", *Management Science*, 55(7), 1077-1093.

OCDE (2013), "Exploring the Economics of Personal Data : A Survey of Methodologies for Measuring Monetary Value ", *OECD Digital Economy Papers*, N° 220, Éditions OCDE, Paris.

TAYLOR C. A. (2004), "Consumer privacy and the market for customer information", *RAND Journal of Economics*, 35(4), 631-665.

The Economist (2010), "Data, data everywhere", *The Economist Newspaper Limited*.

TUCKER C. E. (2014), "Social Networks, Personalized Advertising, and Privacy Controls", *Journal of Marketing Research*, 51(5), 546-562.

TUCKER C. E. (2012), "The Economics of Advertising and Privacy", *International Journal of Industrial Organization*, 30(3), pp. 326-329.

VARIAN H. R. (1997), "Economic aspects of personal privacy", *Privacy and Self-Regulation in the Information Age*, US Department of Commerce.

La donnée, une marchandise comme les autres ?

Par Henri ISAAC
PSL Université Paris-Dauphine, Dauphine Recherches Management

Avec la numérisation du monde, sa mise en données (« *datafication* ») n'a eu de cesse de s'étendre au point de désormais concerner à peu près tous les aspects de l'activité humaine. Cette prolifération des données génère une nouvelle économie dans laquelle des acteurs économiques s'emparent de cette nouvelle matière spécifique qu'est la donnée pour produire des services numériques toujours plus nombreux. La donnée apparaît donc comme la nouvelle marchandise de ce XXI^{e} siècle. Sa génération, sa captation, sa possession et son exploitation sont perçues comme la source de création de valeur. Dès lors, nombreux sont ceux qui considèrent la donnée comme une marchandise, un bien échangeable, source de richesse. Plus encore, des sociétés proposent aux particuliers de commercialiser leurs données personnelles, à l'instar de Datacoup [(1)] (Elvy, 2017), ou d'autres encore organisent une place de marché des données (Dawex [(2)]) sur laquelle les entreprises peuvent commercialiser des jeux de données. D'autres tentent de monétiser les données de consommateurs directement auprès des annonceurs (Wysker [(3)]).

La donnée numérique apparaît donc désormais aux yeux de nombreux acteurs économiques comme la matière première d'une économie numérisée, ce que les *data brokers* avaient déjà largement compris dès le milieu des années 1970 [(4)], au début de l'ère du marketing direct. Cependant, les caractéristiques de la donnée sont loin d'en faire une marchandise comme les autres. En effet, un examen minutieux des caractéristiques de la donnée met en évidence que sa production dans l'ère digitale est souvent un mécanisme complexe qui l'éloigne de la notion de marchandise (Isaac, 2018), et que sa valeur d'usage s'avère bien plus complexe à définir que dans le cas de la marchandise. Certains vont même jusqu'à considérer que la production de données numériques s'apparente non pas à une marchandise mais à une forme de travail (Arrieta Ibarra *et al.,* 2017). En outre, les données sont l'objet de différentes régulations juridiques, particulièrement en Europe, où plusieurs règles juridiques (données personnelles, *open data*) les éloignent définitivement de la notion de marchandise.

De la donnée : valeur d'usage, valeur d'échange de la donnée

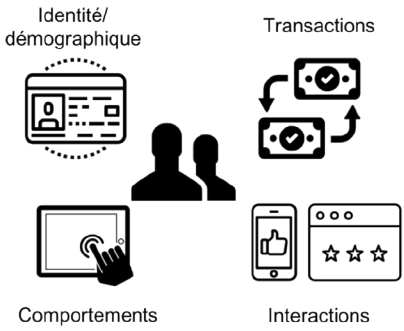
Considérer la donnée comme une marchandise nécessite d'éliciter la notion de marchandise que l'on retient. Plusieurs définitions de la marchandise existent. Deleplace parle de marchandise dès que « *quelque chose apparaît comme objet de transaction au sein d'un marché, composé d'une triade d'éléments caractéristiques : l'offreur, le demandeur, le prix. Position minimaliste : est marchandise "ce qui est transféré d'un individu à un autre en échange de la monnaie qu'il reçoit*" ». Cette définition, tout comme celle que Polanyi propose (« *Les marchandises sont ici empiriquement définies comme des objets produits pour la vente* »), permettent de comprendre assez aisément que de nombreuses

(1) https://datacoup.com
(2) https://www.dawex.com
(3) https://www.wysker.com
(4) Voir le rapport *Data Brokers: A Call For Transparency and Accountability* de la FTC https://www.ftc.gov/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014

données ne constituent pas des marchandises. En effet, dans l'espace numérique, nombre d'entre elles sont produites sans que la finalité de les vendre existe, puisque nombre d'entre elles sont produites sans que les personnes qui les produisent en aient elles-mêmes conscience. Ainsi, les données issues des recherches d'un internaute dans le moteur de recherche d'un site marchand ne constituent pas une donnée pour l'auteur de ces recherches. En revanche, le traitement secondaire de la recherche en fait une donnée utile pour faire apparaître les intentions de consommation sur le site. La valeur est donc produite postérieurement à l'échange et la valeur d'usage n'est définie ni dans l'échange, ni par son producteur.

À cet égard, l'approche marxiste de la marchandise est utile pour questionner le statut de la donnée comme valeur à la fois d'usage et d'échange. En effet, selon Marx, la marchandise se définit lorsqu'un bien est à la fois valeur d'échange et valeur d'usage. La valeur d'usage est incluse dans le bien et elle préexiste à l'utilisation du bien. Pour Marx, l'homme ne crée pas de la valeur d'usage, elle est tout entière dans le bien ; l'activité de l'homme consiste à découvrir les propriétés du bien et à les rendre plus accessibles. En aucun cas le travail ne crée de la valeur d'usage, il la révèle. Marx écrit : « L'utilité d'une chose fait de cette chose une valeur d'usage. Mais cette utilité n'a rien de vague et d'indécis, déterminée par les propriétés du corps de la marchandise, elle n'existe point sans lui... Ce corps lui-même, […], est conséquemment une valeur d'usage, et ce n'est pas le plus ou moins de travail qu'il faut à l'homme pour s'approprier les qualités utiles qui lui donne ce caractère [5] ». La valeur d'échange est déterminée par la quantité de travail incorporée dans un bien. C'est donc dans la sphère de production que la valeur d'échange est déterminée. Mais le produit doit également être une valeur d'usage pour autrui, c'est-à-dire que le produit doit comporter une valeur d'usage *sociale*.

Or, dans le cas de la donnée, l'aliénation et l'emploi de la valeur d'usage sont loin d'être simultanés et ne sont pas non plus causés par le producteur de la donnée. C'est précisément une des caractéristiques de la phase actuelle de numérisation que d'ajouter aux données personnelles, données liées à l'identité et aux transactions, des données qui sont produites dans les dispositifs digitaux comme les sites Internet, applications mobiles, caméra, capteurs, objets connectés, etc., et qui ne sont que les traces des navigations et des usages, ce que d'aucuns caractérisent précisément par la notion d'économie de la trace (Kessous, 2011). Ces données sont pour la plupart des données techniques secondaires, des données ancillaires (*by-product computing*), c'est-à-dire qu'elles n'ont, pour celui qui en est à l'origine, aucune valeur directe. Ces données ont, pour une grande partie d'entre elles, une durée de vie utile très limitée dans le temps. Ainsi, les recherches sur un site marchand ou les navigations sur de tels sites caractérisent, grâce à un traitement, des comportements de consommation pour la seule période de navigation sur ces sites. Leur valeur d'usage est donc très limitée dans le temps. Dans certains cas, cette valeur d'usage est encore plus limitée puisqu'elle ne s'exprime que lors d'enchères en temps réel, comme dans le cas de la publicité en ligne achetée grâce aux techniques de l'achat programmatique (Balusseau, 2018).



Extension du champ des données personnelles à l'ère numérique.

Cependant, le point crucial est que la valeur d'usage des données se produit, pour nombre d'entre elles, postérieurement à l'échange dans des traitements qui les agrègent et les anonymisent pour en faire des biens échangeables, notamment sur le marché publicitaire digital sur lequel les différents acteurs mobilisent leurs données et les combinent avec celles des éditeurs ou d'autres fournisseurs.

(5) MARX K., *Le Capital*, Paris, 1872, Livre premier, première section, chapitre premier

En outre, l'analyse des différents types de données que la numérisation des échanges produit met en évidence leur grande variété. Comme l'illustre le tableau, il est possible de distinguer des données qui préexistent à tout échange et dont la valeur d'usage préexiste à celui-ci, leur conférant un véritable statut de marchandise.

	Préexiste à l'échange	Produites dans l'échange
Produites par l'utilisateur	Données d'identité de l'individu	Données transactionnelles Données produites par l'utilisateur
Produites par le dispositif d'échange	Données contractuelles	Données industrielles (M2M) Données de navigation Données comportementales Données ancillaires (*by-product*)[9]

Typologie des régimes de production des données

En revanche, de très nombreuses données ne sont produites que dans l'échange ou postérieurement à son exécution (profil d'achat d'un client, par exemple). La valeur d'usage de la donnée ne se révèle donc que dans le temps. C'est encore plus vrai dans la phase actuelle de l'intelligence artificielle qui repose sur des jeux de données accumulés dans le temps. Ceci est particulièrement saillant dans les algorithmes auto-apprenants (*machine learning*) ou ceux d'apprentissage profond (*deep learning*) qui reposent sur l'accumulation d'un grand nombre de données pour bâtir des modèles de décision dont l'usage s'améliore automatiquement au fur et à mesure qu'ils incorporent de nouvelles données [(7)].

Dès lors, ces données ne peuvent être considérées, selon la définition retenue, comme des marchandises, puisque leur valeur d'usage n'existe parfois même pas au moment de l'échange, ou en tout cas est alors impossible à définir mais se révélera à l'usage.

L'analyse économique montre donc que beaucoup de données ne peuvent être considérées comme une marchandise. Ceci est encore renforcé par le fait que plusieurs régimes juridiques empêchent de considérer un grand nombre de données comme une marchandise.

Régime juridique et valeur de la donnée

Les données relèvent principalement de trois régimes juridiques distincts : le régime des données personnelles, l'*open data*, et la base contractuelle classique régissant l'échange des données ni personnelles ni publiques.

Les deux premiers régimes juridiques considèrent l'un et l'autre, sur des fondements juridiques très différents, que les données ne sont pas des marchandises, que les données personnelles relèvent du droit de la personne et que les données publiques ouvertes relèvent d'un régime juridique spécifique qui ne permet pas de considérer celles-ci comme des marchandises, mais comme un bien commun accessible à tous.

Les données personnelles des individus ne sont en effet pas considérées comme des propriétés, mais comme un droit attaché à la personne humaine. C'est ainsi que la loi Informatique & Libertés de 1978 et le Règlement européen sur la Protection des Données de 2016 les ont conçues.

Dans le cadre européen, l'usager a le droit d'exiger que s'interrompent la collecte et le traitement de ses données et il a aussi désormais un droit de minimisation des traitements, un droit à la portabi-

(6) Voir Chapitre 1, *in* SCHNEIER B. (2015), *Data and Goliath. The Hidden Battles to Collect your Data and Control Your World*
(7) Voir ISAAC H. (2018).

lité de ses données et un droit à l'oubli. Le citoyen reste titulaire du droit fondamental de contrôler ses données personnelles, même s'il en a accordé l'usage à l'exploitant d'un service numérique. Ce dernier n'en est pas le propriétaire, mais bien l'utilisateur sous conditions.

Il n'y a donc pas transfert de propriété et l'échange ne peut être considéré comme marchand [8]. En revanche, les données peuvent faire l'objet d'un échange marchand dès lors que l'utilisateur en a consenti le droit à l'opérateur du service numérique. Par ailleurs, les données personnelles, rendues anonymes, une fois agrégées peuvent faire l'objet de transactions marchandes et devenir alors des marchandises qui font l'objet d'un marché avec les spécialistes que sont les *data brokers*.

Par ailleurs, les informations en elles-mêmes échappent au droit d'auteur et à la propriété intellectuelle. Elles appartiennent par défaut au domaine public structurel, même si l'instauration d'un droit des bases de données à partir des années 1990 en Europe a tendu à les soumettre peu à peu à une logique propriétaire qui ne cesse de progresser depuis lors, élargissant progressivement le champ des bases de données marchandes, comme c'est le cas notamment avec les bases de données scientifiques ou les bases de données économiques. Il faut ici noter que la donnée prise isolément n'a guère de valeur intrinsèque et qu'elle est très loin de constituer une marchandise.

Il en va de même pour les données publiques, désormais soumises à un régime d'ouverture par défaut depuis la loi pour une République numérique de 2016. L'ouverture des données publiques participe d'un mouvement plus large, l'Open Government Partnership (OGP), qui vise à rendre plus transparente l'action publique, notamment par la mise à disposition des données publiques sous un format automatiquement réutilisable (Chignard, 2012).

La licence associée aux données publiques ne permet pas aux administrations de les revendre. Elles sont en revanche susceptibles d'être à la base d'un service marchand de la part du réutilisateur. Il n'est donc pas possible de considérer ces données comme des marchandises.

En revanche subsistent des données, ni publiques ni personnelles, qui ne sont actuellement soumises à aucune réglementation. C'est le cas des données des machines agricoles, des drones, des objets connectés, et plus généralement des données brutes issues des machines (*machine-generated raw-data*).

Les cadres légaux de l'Europe et de la France limitent donc strictement la possibilité pour une donnée de devenir une marchandise dans ces espaces, ce qui n'est pas forcément le cas d'autres espaces juridiques comme les États-Unis, ce qui amène les Européens à négocier avec ce pays un accord (Privacy Shield) pour préserver les garanties attachées aux données personnelles lors des transferts de données dans cette juridiction.

Conclusion

À l'approche de l'entrée en vigueur du nouveau Règlement européen sur la Protection des Données en mai 2018, des voix s'élèvent pour attacher à la donnée personnelle un droit de propriété qui permettrait aux individus d'en faire le commerce, notamment avec les grandes plateformes numériques. Cette approche qui cherche à « patrimonialiser » la donnée personnelle la considère donc comme une marchandise comme les autres.

Elle ignore la nature des données qui sont utilisées par les services numériques, données pour l'essentiel ancillaires par rapport aux prestations, produites en temps réel, et qui n'ont, pour l'utilisateur qui les produit, aucune valeur directe ni par leur propriété, ni par leur commerce. Il serait

(8) Ceci est largement confirmé par le Conseil d'État en 2014 et par le Conseil national du Numérique en 2017.

donc possible, dans bien des cas, de considérer les données personnelles et les données publiques ouvertes comme une « marchandise fictive » au sens de Polanyi [9].

Références

ALARY J., BALUSSEAU V. (2018), *La Publicité à l'heure de la data : Ad tech et programmatique expliqués par des experts,* 276 p., Dunod.

ARRIETA IBARRA I., GOFF L., JIMÉNEZ HERNÁNDEZ D., LANIER J. & WEYL E. G. (2017), "Should We Treat Data as Labor? Moving Beyond 'Free'", American Economic Association Papers & Proceedings, Vol. 1, n° 1, https://ssrn.com/abstract=3093683

AZAM G. (2007), « La connaissance, une marchandise fictive », *Revue du MAUSS*, n° 29, pp. 110-126. DOI 10.3917/rdm.029.0110

CHIGNARD S. (2012), *Open Data*, fyp editions, 191 p.

Conseil d'État (2014), *Le Numérique et les Droits fondamentaux*, La Documentation française, 446 p.

Conseil national du Numérique (2017), *La Libre Circulation des données dans l'Union européenne.*

DELEPLACE G. (1979), *Théorie du capitalisme : une introduction*, PUG-Maspero.

DUCH-BROWN N., MARTENS B. & MUELLER-LANGER F. (2017), "The Economics of Ownership, Access and Trade in Digital Data", *JRC Digital Economy Working Paper,* 2017-01, https://ssrn.com/abstract=2914144

ELVY S.A. (2017), "Paying for Privacy and the Personal Data Economy", *Columbia Law Review*, Vol. 117, n° 6, october, https://ssrn.com/abstract=3058835

ISAAC H. (2018), « La donnée numérique, bien public ou instrument de profit », *Pouvoirs,* n° 164, numéro spécial « La Datacratie », pp. 75-86, janvier.

KESSOUS E. (2011), « L'économie de l'attention et le marketing des traces », Actes du colloque « Web social, communautés virtuelles et consommation », 79e congrès international ACFAS, Chaire de relations publiques et communication marketing – UQAM, Université de Sherbrooke, mai 2011.

MANNE G. A. & SPERRY B. (2015), "The Law and Economics of Data and Privacy in Antitrust Analysis", 2014 TPRC Conference Paper, https://ssrn.com/abstract=2418779

POLANYI K. (1983), *La Grande Transformation. Aux origines politiques et économiques de notre temps*, Paris, Gallimard.

SCHNEIER B. (2015), *Data and Goliath. The Hidden Battles to Collect your Data and Control Your World*, W.W. Norton & Company, 383 p.

(9) Voir AZAM G., 2007.

Données personnelles et éthique : les enjeux économiques de la confiance

Par Patrick WAELBROECK [1]
Professeur d'économie, Télécom ParisTech

Nous laissons de nombreuses traces lorsque nous utilisons l'Internet ou d'autres outils numériques. Ces traces sont stockées et analysées par les moteurs de recherche et navigateurs du web. Si certains internautes laissent ces traces de manière plutôt involontaire, d'autres contribuent de manière active aux communautés socio-numériques (comme celles d'eBay, d'Amazon, de Wikipédia, de Twitter, de YouTube) en laissant des notes, des avis, des commentaires, des classements de produits et services, ou en publiant des fichiers. Par leurs traces et contributions, les internautes et les utilisateurs d'outils numériques sont des producteurs d'informations personnelles. L'économie numérique exploite ces traces et contributions pour construire ses modèles d'affaires. Les sites financés par la publicité ciblent, voire reciblent, les prospects afin de leur fournir des contenus et des prix personnalisés. De plus, les entreprises du Net telles qu'Amazon ou Netflix, pour développer leurs modèles commerciaux, utilisent ces données personnelles afin d'identifier les préférences des internautes et de leur faire des recommandations personnalisées. De même, eBay utilise les notes laissées par ses utilisateurs sur leurs transactions pour construire un système de réputation en ligne. YouTube et Facebook doivent leur existence au contenu généré par leurs utilisateurs.

Si l'économie numérique se nourrit des données personnelles fournies parfois volontairement par les internautes, il n'en demeure pas moins que de plus en plus d'utilisateurs de réseaux sociaux sont préoccupés par la manière dont leurs données sont exploitées. En effet, on compte désormais par millions les vols de données des clients de compagnies telles que Yahoo, Equifax, eBay, Sony, LinkedIn, ou encore plus récemment Cambridge Analytica. Les révélations de l'affaire Snowden sur la surveillance généralisée par les États ont également créé un sentiment de défiance envers les acteurs de l'économie numérique à tel point qu'en 2015, 21 % des internautes n'étaient prêts à partager aucune information ; ils n'étaient que 5 % en 2009 [2]. N'y a-t-il pas un coût sociétal au tout gratuit sur Internet ?

Le numérique bouleverse les conditions de l'échange à travers l'asymétrie d'information qu'il engendre, dans ce que F. Pasquale appelle la « black box society » : les utilisateurs d'outils numériques ne connaissent ni l'utilisation qui est faite de leurs données personnelles, ni le volume des données échangées par les entreprises qui les collectent. Pire encore, ces entreprises peuvent manipuler le contexte informationnel de la transaction pour mettre les individus dans un environnement qu'ils pensent être de confiance (Mantelero, 2013) afin de les inciter à divulguer plus d'informations personnelles.

Cet article cherche à apporter quelques éléments d'éclaircissement sur les enjeux économiques de la confiance, qui peut être appréhendée d'un point de vue économique par les risques liés à une transaction. Ces risques peuvent porter sur les termes d'une transaction individuelle ou sur l'environnement institutionnel dans lequel se déroule cette transaction.

(1) L'auteur remercie les membres de la chaire Valeurs et politiques des informations personnelles pour les discussions stimulantes ayant contribué à l'enrichissement de cet article.
(2) Baromètre de la confiance de l'ACSEL-CDC.

Comprendre la source économique des risques

Il s'agit de savoir tout d'abord quels types de données sont utilisés par les entreprises et quels sont les risques pour les consommateurs. La collecte de données personnelles crée des risques dans le cas de traces involontaires, laissées par un internaute dans son historique de navigation, car l'utilisateur n'a pas toujours conscience des conséquences des traitements effectués par les entreprises du numérique. Dans le cas de contributions volontaires, comme celle d'un consommateur commentant un blog ou évaluant la qualité d'un produit ou la réputation d'un vendeur, il existe toujours des risques liés au vol de données. Les données utilisées à mauvais escient peuvent conduire à des externalités négatives telles que le vol d'identité, le harcèlement en ligne et la divulgation d'informations intimes, l'utilisation frauduleuse de coordonnées bancaires, la discrimination par les prix, les publicités indésirables, ciblées ou non.

Les externalités négatives résultent d'une défaillance du marché lorsque les actions d'un agent économique exercent un effet négatif sur d'autres agents sans compensation liée à un mécanisme de marché. Ces externalités, négatives pour le consommateur, sont causées par des entreprises qui collectent trop de données par rapport à l'optimum social, et leur existence justifie économiquement la protection des données personnelles.

Le problème de l'utilisation abusive des données personnelles est renforcé par l'asymétrie d'information en ligne. D'une part, il est difficile pour un consommateur de vérifier comment ses données sont utilisées par les compagnies qui les collectent et les traitent, et encore plus de savoir si cette utilisation est conforme ou non à la législation. Ceci est encore plus vrai à l'ère du Big Data où des bases de données indépendantes contenant peu d'informations personnelles peuvent être combinées facilement pour identifier une personne. D'autre part, un individu est difficilement capable d'évaluer techniquement le niveau de sécurité informatique dont font l'objet ses données pendant leur transmission et leur stockage. Parfois, les entreprises elles-mêmes ne sont pas toujours en mesure d'évaluer totalement la sécurité de leur système d'information : elles ignorent parfois si elles ont subi une cyberattaque. Les enjeux économiques sont de taille. Dans les travaux qui lui ont valu le prix Nobel d'économie, Akerlof a montré que des situations d'asymétrie d'information pouvaient conduire à la disparition de marchés. L'économie numérique n'échappe pas à cette théorie, et la détérioration de la confiance crée des risques qui peuvent conduire certains internautes à se « débrancher » [(3)].

On peut distinguer deux types de risques. Le premier est lié à la transaction individuelle. Il s'agit essentiellement de risques idiosyncratiques de contrepartie. Le contrat va-t-il être respecté ? Tous les termes du contrat seront-ils appliqués ? Ces risques ne sont pas toujours assurables car il est difficile de formuler toutes les éventualités du contexte du contrat, une situation que les économistes qualifient de contrat incomplet. Le deuxième type de risques est lié à l'environnement dans lequel le contrat s'exécute. Ces risques sont de nature systémique. Existe-t-il un recours en cas de problème ? Il est évident que la transaction s'inscrit dans un cadre juridique. Néanmoins les bénéfices d'une procédure légale ne couvrent souvent pas les coûts de transaction pour de faibles montants. Illustrons ces différents points dans le contexte d'une transaction eBay. Un acheteur commande un produit. Sera-t-il livré à temps ? Le produit correspond-il bien à celui commandé ? Il existe bien un risque idiosyncratique lié à la transaction. Il est atténué par un environnement de confiance dans lequel se déroule la transaction. En effet, le risque systémique peut être réduit à travers deux mécanismes. Le premier est un mécanisme punitif : l'acheteur trompé et déçu peut donner une mauvaise évaluation au vendeur. Le deuxième mécanisme est lié au tiers de confiance Paypal qui permet d'entamer une procédure de litige qui peut aboutir à un remboursement.

(3) Étude sur les données personnelles, chaire Valeurs et politiques des informations personnelles, 2017, https://cvpip.wp.imt.fr/donnees-personnelles-et-confiance-quelles-strategies-pour-les-citoyens-consommateurs-en-2017/

Enjeux économiques des problèmes d'éthique soulevés par les stratégies numériques

En plus des externalités négatives évoquées précédemment, l'utilisation des données personnelles dans les stratégies commerciales des acteurs du numérique soulève des considérations éthiques qui ont des conséquences économiques de premier ordre. Nous en analysons trois : le libre arbitre, ou possibilité de choisir dans un environnement informationnel neutre ; l'autonomie, ou capacité à se conformer à ce qui est bon pour soi ; la discrimination, ou absence de biais systématique.

Libre arbitre et autonomie

Pour pouvoir prendre une décision économique optimale, les consommateurs doivent pouvoir choisir dans un environnement informationnel neutre. Deux phénomènes, exacerbés par les stratégies de l'économie numérique évoquées précédemment, posent problème.

Premièrement, les consommateurs reçoivent des informations filtrées par les plateformes telles que Google ou Facebook. Par exemple, le moteur de recherche Google filtre les résultats de recherche en fonction de la géolocalisation, de l'historique de navigation et du profil publicitaire. Facebook filtre les informations en fonction des « likes » et du profil de l'utilisateur. Ces filtres d'information peuvent influencer le comportement des internautes. Ils soulèvent des problèmes économiques importants liés principalement à la construction des préférences. En effet, les bulles informationnelles sont créées par des algorithmes qui génèrent un univers spécifique à un internaute, univers qui peut potentiellement influencer la manière dont il pense, se comporte et achète.

Deuxièmement, Amazon, Netflix et Spotify, parmi beaucoup d'autres entreprises du numérique, exécutent des algorithmes pour fournir des recommandations de produits personnalisés en fonction de l'historique de navigation et des achats d'une personne.

Si l'environnement informationnel peut être manipulé, les agents économiques ne prennent plus les décisions optimales. Les questions du libre arbitre et de l'autonomie se traduisent par des choix économiques qui peuvent être manipulés par des filtres informationnels et des recommandations ciblées.

Discrimination et biais

La collecte de données personnelles permet également de pratiquer des stratégies de discrimination par les prix, à savoir vendre le même produit ou service à différents prix nets à des consommateurs différents. Le prix net comprend les frais de livraison et de production. Les entreprises cherchent le meilleur prix auquel elles peuvent vendre leurs produits et services en fonction de la disponibilité à payer des clients potentiels. Pour les produits numériques, la forme de discrimination la plus répandue consiste à développer des stratégies pour identifier plusieurs groupes de consommateurs et proposer différentes versions d'un même produit ou service à ces groupes. Par exemple, un fabricant de logiciels propose un même produit avec différentes fonctionnalités : une version professionnelle complète et une version basique, ou étudiante, pour laquelle certaines fonctions ne sont pas disponibles. Les informations personnelles des consommateurs peuvent donc être utilisées pour personnaliser les offres à des clients ciblés, souvent à un coût très faible. Certains consommateurs bénéficient de prix bas, mais d'autres se voient proposer le produit à des prix plus élevés et peuvent décider de protéger leurs données personnelles pour éviter d'être discriminés.

Conséquences économiques

La manipulation de l'environnement informationnel et le ciblage posent des questions éthiques qui ont de réelles conséquences économiques, en particulier du point de vue de la concurrence et de l'innovation sur les marchés. Nous considérons par la suite cinq points préoccupants.

Premièrement, les algorithmes qui influencent les choix des consommateurs par l'imposition de valeurs externes modifient la manière dont les consommateurs construisent leur fonction d'utilité. On peut penser à Facebook qui censure des œuvres d'art jugées choquantes (*L'Origine du monde* par exemple) ou qui nourrit des algorithmes avec des données collectées sur des citoyens américains pour les appliquer à leurs utilisateurs à travers le monde. Les problématiques de la confiance deviennent alors des enjeux culturels. Deuxièmement, en l'absence de concurrence forte, les plateformes qui contrôlent l'accès des utilisateurs aux données peuvent développer des stratégies de forclusion ou refuser de faire commerce avec des entreprises tierces. Il existe de nombreux exemples de cette pratique. En juin 2012, Facebook imposait comme adresse de messagerie par défaut l'adresse du domaine @facebook.com. Toujours en juin 2012, Apple annonçait qu'il installait par défaut son logiciel de cartographie Plans à la place du logiciel Google Maps dans son nouveau système d'exploitation iOS. En avril 2013, Apple retirait de son App Store l'application AppGratis qui faisait chaque jour la promotion d'une application devenue temporairement gratuite sur le Store. En 2001, les États-Unis demandaient à Microsoft de supprimer l'offre groupée du système d'exploitation Windows et du navigateur Internet Explorer pour laisser au consommateur la possibilité d'installer le navigateur de son choix, et d'ouvrir l'interface de programmation à des compagnies tierces. En juin 2017, la Commission européenne inflige une amende de 2,4 milliards d'euros à Google pour avoir filtré des résultats de recherche en mettant en avant Google Shopping tout en défavorisant les concurrents. Troisièmement, qui garantira que les outils de protection de la vie privée seront disponibles pour les consommateurs ? Ces solutions techniques sont coûteuses à développer et vont à l'encontre des stratégies du numérique et de surveillance des États. De manière générale, la sécurité informatique est un bien public qui profite à tout le monde. Il existe donc un risque de sous-investissement des entreprises en protection des données personnelles qui peut conduire à davantage de fuites de données et de perte de confiance [(4).] Quatrièmement, les plateformes qui contrôlent l'accès aux données des internautes augmentent les coûts d'entrée sur le marché et peuvent freiner l'innovation qui nécessite des données personnelles. Cinquièmement, les algorithmes qui personnalisent les prix en fonction des consommateurs peuvent être détournés pour faciliter la collusion entre entreprises [(5)]. À titre d'illustration, on peut citer l'exemple du livre qui coûtait 23 millions de dollars, à la suite d'une surenchère algorithmique entre deux vendeurs désireux d'optimiser leur profit sur la plateforme Amazon Marketplace [(6)].

Conclusions et pistes de réflexion

La confiance repose sur l'évaluation des risques encourus lors d'une transaction effectuée dans l'économie numérique. Nous avons classé ces risques en deux catégories : idiosyncratiques et systémiques. Il est donc nécessaire pour développer la confiance dans l'économie numérique d'agir sur ces deux leviers. Premièrement, il faut renforcer la connaissance des internautes sur la manière dont les données sont utilisées et sur les externalités négatives. Deuxièmement, il s'agit de rétablir un sentiment de réciprocité et d'équité en garantissant un échange de valeur équitable (l'échange d'un service gratuit contre des données personnelles ne semble plus satisfaisant) et un mécanisme punitif crédible en cas d'utilisation abusive des données personnelles. La sanction renforcée du Règlement général sur la Protection des Données (RGPD) entrant en vigueur en mai 2018 va dans ce sens.

(4) Lire Dubus et Waelbroeck (2018) à ce sujet.
(5) Lire le rapport de l'OCDE (2017).
(6) Lire http://edition.cnn.com/2011/TECH/web/04/25/amazon.price.algorithm/index.html
Il n'y avait que deux vendeurs, des robots, dont l'un possédait le livre à vendre et voulait se situer juste au-dessous du prix demandé par l'autre, qui, lui, ne possédait pas le livre et demandait un prix sensiblement supérieur à celui du premier dans l'intention de lui acheter le livre et de réaliser un beau profit à la revente.

L'appropriation par les consommateurs des nouveaux outils de protection forme un autre moyen de construire la confiance. Des logiciels permettant de masquer les adresses IP, les extensions de navigateur Internet bloquant les scripts et les publicités rendent plus difficile l'identification des utilisateurs et la discrimination par les prix. Nous avons montré dans une enquête réalisée en 2017 sur un échantillon représentatif de la population française que ceux qui se protègent le plus sont ceux également qui achètent le plus sur Internet [7]. Ce résultat peut sembler paradoxal de premier abord, mais devient logique lorsqu'on comprend que ces outils permettent aux consommateurs d'acheter en toute confiance. On ne peut plus considérer les consommateurs comme passifs face aux stratégies des acteurs du numérique.

La question de la confiance dans l'utilisation des données porte également sur la valeur économique de l'anonymat. Une théorie simpliste postule qu'il existe un arbitrage entre valeur économique d'une part et protection de la vie privée et anonymat d'autre part. Il y aurait ainsi deux situations extrêmes : une situation où une personne est parfaitement identifiée et susceptible de recevoir des offres ciblées et une autre situation où la personne serait anonyme. Dans le premier cas, la valeur économique serait maximale dans le second ; les données n'auraient pas de valeur. Si l'on déplace le curseur vers le ciblage, on augmente la valeur économique au détriment de la protection de la vie privée. Inversement, si l'on protège la vie privée, on réduit la valeur économique des données. Cette théorie ignore les enjeux de confiance dans l'utilisation des données. Pour développer une relation client sur le long terme, on doit raisonner en termes de risques et d'externalités pour le client. Il existe donc une valeur économique à la protection de la vie privée, qui tourne autour de la relation de long terme et de la notion de confiance, et de la garantie du libre arbitre, de l'autonomie et de l'absence de discrimination. Les principes de pseudonymisation et de consentement explicite du RGPD vont dans ce sens.

Bibliographie

DUBUS A. et WAELBROECK P. (2018), «La notion de confiance en économie,» in : *Signes de confiance – L'impact des labels sur la gestion des données personnelles*, Levallois-Barth (Ed.), https://cvpip.wp.imt.fr/2018/01/30/8-03-2018-signes-de-confiance-limpact-des-labels-sur-la-gestion-des-donnees-personnelles/

MANTELERO A. (2013). "Competitive value of data protection: the impact of data protection regulation on online behavior", *International Data Privacy Law* 3(4): 229-238. DOI : 10.1093/idpl/ipt016

OCDE (2017), *Algorithms and Collusion: Competition Policy in the Digital Age*, www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm

(7) Étude sur les données personnelles, chaire Valeurs et politiques des informations personnelles, 2017, https://cvpip.wp.imt.fr/donnees-personnelles-et-confiance-quelles-strategies-pour-les-citoyens-consommateurs-en-2017/

Les sources d'inspiration du Règlement général sur la Protection des Données : la conformité, la réglementation de l'environnement, la responsabilité du fait des produits défectueux

Par Winston MAXWELL et Christine GATEAU
Avocats associés, Hogan Lovells

Le Règlement général sur la Protection des Données à caractère personnel (RGPD) s'appuie sur les mêmes principes que la directive 95/46, principes qu'on retrouve aussi bien dans la Convention 108 du Conseil de l'Europe de 1981 que dans les lignes directrices de l'OCDE de 1980, révisées en 2013. Ces mêmes principes se retrouvent dans la loi Informatique et Libertés (LIL) de 1978 et le *Privacy Act* américain de 1974. Les principes fondamentaux n'ont pas fondamentalement changé depuis quarante ans. Les données doivent être traitées de manière loyale et licite, collectées pour des finalités déterminées, légitimes et non excessives, elles doivent être exactes, et conservées pendant une durée n'excédant pas celle nécessaire aux finalités pour lesquelles elles ont été enregistrées.

Le RGPD représente néanmoins une rupture avec le passé dans la mesure où il prévoit un système de responsabilisation des entreprises et un régime de sanctions dissuasives, à un niveau inédit en matière de données à caractère personnel. Ce qui a changé avec le RGPD, c'est l'ampleur des risques − et des opportunités − liés à l'exploitation des données. Pour rester efficace, la régulation a dû changer d'échelle. Les données deviennent un sujet de conformité majeur, à l'instar des règles anti-corruption, du droit de la concurrence ou de la réglementation de l'environnement et des installations dangereuses. Le RGPD représente une prise de conscience similaire à celle qui s'est opérée à la fin des années 1970 et au début des années 1980 après le naufrage de l'***Amoco Cadiz*** en 1978 et la catastrophe de Seveso en 1976. Le RGPD s'inspire d'ailleurs de la réglementation des installations SEVESO puisque les mesures de traçabilité, de stockage et d'utilisation des données seront adaptées selon leur niveau de « toxicité » potentielle. L'exploitant doit identifier les risques à leur source et élaborer lui-même le plan de sécurité pour l'installation, sous la supervision du régulateur.

Les données à caractère personnel sont devenues indispensables à l'économie mais, comme le pétrole et d'autres produits chimiques, elles peuvent créer des externalités négatives importantes, nécessitant une intervention réglementaire musclée.

En Europe, les données à caractère personnel sont à la fois un objet de commerce et un droit fondamental. Le RGPD tente de réconcilier ces deux principes. L'objectif du règlement est de faciliter la libre circulation des données à caractère personnel et leur exploitation par les entreprises. En même temps, le RGPD nous rappelle que les données ne sont pas des services et marchandises comme les autres. Le règlement tente de gérer cette tension, ainsi que la tension qui existe entre différents droits fondamentaux eux-mêmes. Une mesure qui augmente la protection de la donnée à caractère personnel peut nuire au droit à l'accès à l'information ou au droit à la protection de la propriété privée. En cas de frottement entre plusieurs droits et libertés fondamentaux, une règle de proportionnalité s'applique afin d'assurer la limitation de chaque interférence au strict nécessaire.

Une direction : le principe d'*accountability*

Le RGPD oblige l'entreprise à mettre en place un cadre pour effectuer elle-même des analyses de risques et arriver à ses propres conclusions quant au caractère loyal, non excessif et adéquat des traitements et des mesures d'accompagnement. Ce cadre interne repose d'abord sur la création d'un registre pour chaque traitement de données à caractère personnel. Le registre oblige l'entreprise à répertorier chaque action de traitement, à identifier l'entité responsable du traitement, l'existence de sous-traitants et de transferts internationaux. Surtout, le registre contient une description précise de la finalité du traitement : pourquoi est-ce que je traite ces données ? Le registre est la clé de voûte du système de responsabilisation. Il permettra de mettre en évidence le traitement de données à risques. Il permettra d'identifier des traitements pour lesquels les finalités sont vagues ou mal définies. Il permettra d'identifier des traitements faisant appel à des sous-traitements ou à des transferts internationaux non encadrés.

Le registre impose une logique de traçabilité, comme pour la gestion de produits dangereux dans une usine. L'absence de registre ou un registre incomplet constitueront automatiquement des fautes aux yeux du régulateur. De même, si le registre contient un traitement à risque et si le responsable du traitement ne fait rien pour réduire ce risque, la faute sera là encore manifeste. Le registre a le mérite de forcer l'entreprise à se poser les bonnes questions : quel est le fondement juridique de ce traitement ? qu'en est-il du consentement des individus ou de l'intérêt légitime de l'entreprise ? le traitement résulte-t-il de l'exécution d'un contrat ? est-ce que la finalité en est suffisamment précise et légitime ? est-ce que les données collectées sont en adéquation avec la finalité ? est-ce que les individus ont reçu une information complète sur le traitement ? comment l'entreprise assure-t-elle l'exercice par les individus de leurs droits d'accès, de rectification, d'effacement et de portabilité ? quelles sont les mesures de sécurité mises en place ? comment avons-nous organisé les transferts de données au sein du groupe ainsi qu'avec nos partenaires et sous-traitants ?

À partir du registre et de cette liste de questions, l'entreprise effectuera un premier bilan des risques et des mesures de conformité mises en place pour les atténuer. Ce premier bilan permettra à l'entreprise de démontrer sa conformité pour l'ensemble des traitements à risque faible ou moyen. Lorsque le premier bilan conduit au constat que le traitement présente un risque élevé, il faut passer à une analyse d'impact détaillée.

Comme la directive SEVESO 2, le RGPD conduit les entreprises et les régulateurs à se concentrer sur les traitements à risques élevés. L'étude d'impact pour les traitements à risque élevé sera un document d'importance capitale pour prouver la conformité. Il s'agit d'un renversement de la charge de la preuve. Dorénavant, il incombe à l'entreprise de démontrer qu'elle a mis en œuvre toutes les mesures appropriées pour protéger les données à caractère personnel en sa possession et assurer le respect des droits des personnes.

Une boussole dans la détermination des « mesures appropriées » : le principe de proportionnalité

L'application du RGPD est un exercice d'équilibriste. La mise en balance de droits et intérêts concurrents, voire contradictoires, se trouve au cœur des concepts-clés du RGPD tels que ceux de « mesures techniques et organisationnelles appropriées », d'« intérêt légitime » et de « traitement loyal ». Ces concepts laissent une grande marge de manœuvre à l'entreprise, aux régulateurs et aux juges pour placer le curseur à différents endroits en fonction du contexte et des risques pour les individus. L'interprétation de ces termes sera différente entre une PME qui gère une base de données de quelques centaines de clients et un géant de l'Internet qui gère les données de dizaines de millions de consommateurs.

Ces multiples zones de frottement ont conduit le législateur à adopter une approche à géométrie variable, fondée sur les risques. Hormis certains cas de figure, il n'existe pas de réponse binaire et univoque dans le règlement. Il pose le principe de « mesures appropriées », concept souple similaire au concept de « bon père de famille » figurant anciennement dans le Code civil et dorénavant remplacé par le terme « raisonnable ». Le règlement met l'accent sur les moyens mis en œuvre par l'entreprise pour assurer une gestion responsable des données à caractère personnel. L'accent mis sur les moyens organisationnels et techniques de protection vient du monde de la conformité ou *compliance*. Les autorités américaines exigent la mise en œuvre de politiques de conformité efficaces en matière de lutte contre la corruption et de droit de la concurrence. Le RGPD s'inspire directement de cette tradition.

L'entreprise doit mettre en œuvre des moyens appropriés, compte tenu des risques pour les individus, de l'état de l'art et des coûts de mise en œuvre. Il s'agit d'une protection raisonnable et non d'une protection absolue.

Mais qu'est-ce qu'une protection raisonnable ? L'analyse d'impact prévue par l'article 35 du RGPD ressemble aux analyses de risques qu'effectuent les entreprises en matière de sécurité des produits avant leur mise sur le marché. En termes d'analyses économiques, des mesures raisonnables doivent correspondre au point où le bien-être social est maximisé. Des mesures de protection trop draconiennes peuvent conduire à un appauvrissement de la société. Par exemple, une régulation qui limite la vitesse des voitures à 30 km/heure réduirait le nombre de victimes d'accidents mais réduirait fortement l'utilité de la voiture. De même, une régulation qui rendrait les plateformes responsables de tous les contenus mis en ligne par les utilisateurs conduirait les plateformes à limiter drastiquement les informations mises en ligne, ce qui conduirait à un appauvrissement de la liberté d'expression. Chaque régulation et chaque mesure de protection peuvent ainsi créer des effets secondaires qui doivent être pris en compte pour déterminer le niveau optimal de régulation.

En termes économiques, le niveau approprié des mesures de protection correspond au point où le coût marginal d'une unité supplémentaire de mesures de protection est égal au bénéfice marginal résultant de cette mesure [(1)]. Au-delà de ce point, les mesures de sécurité supplémentaires coûtent plus cher à la société que le bénéfice qui en découle. Elles vont réduire le bien-être social dans son ensemble, au lieu de l'augmenter.

Le niveau optimal des mesures de protection se situe au point C* dans la figure suivante :

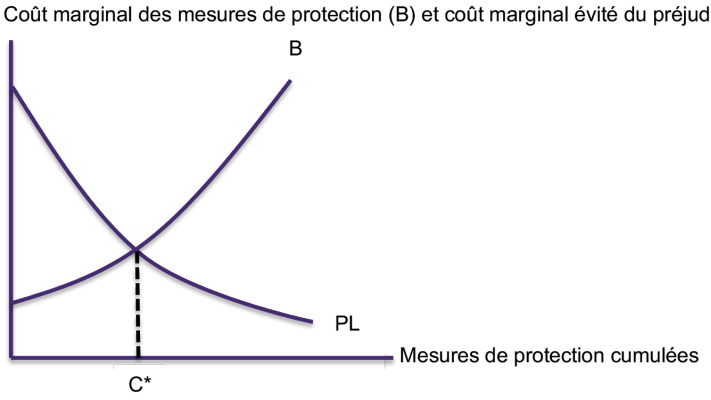


Figure 1 : Illustration de la règle de Hand.
Source : Richard Posner, Economic Analysis of Law, Aspen Casebook Series, 8th Edition, 2011.

(1) Cette règle découle d'une décision de justice américaine rendue par le magistrat Learned Hand en 1947. Depuis lors, cette règle s'appelle la « Règle de Hand ». Elle est utilisée pour définir un comportement fautif.

Dans ce diagramme, la courbe PL représente les coûts liés au risque du préjudice, P étant la probabilité de l'occurrence du préjudice et L étant le niveau de préjudice qui résulterait si le risque se réalisait. Par exemple, si le coût social (L) lié à la perte d'un million de numéros de cartes de crédit est égal à 100 millions d'euros (soit 100 euros par carte) et la probabilité de cette perte (P) est de 0,1 % (une chance sur mille), le produit «PL» est égal à 100 000 euros. Cette courbe PL décroît lorsqu'on ajoute des mesures de protection mais n'atteint pas zéro. La courbe s'aplatit, ce qui signifie qu'au-delà d'un certain seuil, chaque mesure de protection supplémentaire contribue faiblement à la diminution du risque.

La courbe B représente les coûts liés aux mesures de protection. Généralement, les premières mesures de protection, peu onéreuses et très efficaces, ont un impact important sur le risque (diminution de la courbe PL). Mais au-delà d'un certain seuil, les mesures de protection deviennent très chères pour un impact plus faible sur le risque. Par exemple, une mesure de protection qui réduit la probabilité « P » de 100 % à 0,5 % pourrait coûter le même montant qu'une mesure supplémentaire qui réduirait la probabilité « P » de 0,5 % à 0,1 %, ou de 0,1 % à 0,08 %. Chaque incrément de protection devient plus cher lorsqu'on approche un niveau de risque zéro. La courbe B grimpe de manière exponentielle.

Une autre façon de présenter le niveau optimal de mesures de protection est un graphique montrant le point où le bien-être social atteint son niveau maximum :

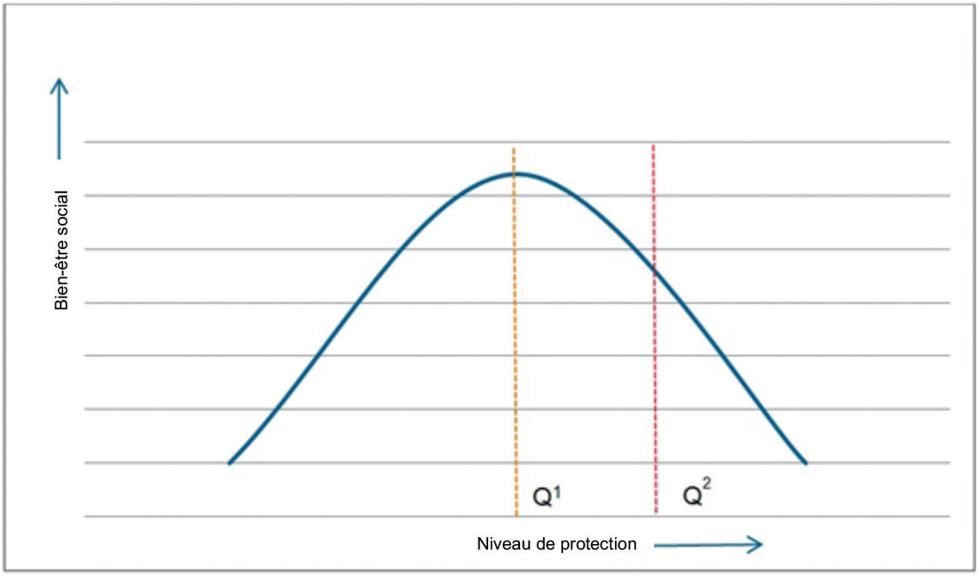


Figure 2 : Maximisation du bien-être social selon le niveau d'efficacité des mesures de protection (Winston Maxwell, *Smarter Internet Regulation Through Cost-Benefit Analysis*, Presses des Mines, 2017).

Même si la mesure de prévention Q2 fournit un niveau de protection supérieur à la mesure Q1, le niveau optimal se trouve au point Q1, car le bien-être social trouve son maximum à ce point.

L'analyse d'impact prévue par l'article 35 du RGPD permettra de mettre en lumière les risques ainsi que les différentes mesures qui peuvent être mises en œuvre pour réduire ces risques. Il appartiendra à la direction de la société de décider quel niveau de risque résiduel est acceptable ou non. Cela dépendra de la culture de l'entreprise et des pratiques de l'industrie dans laquelle l'entreprise évolue. En cas, par exemple, d'une fuite de données, il incombera ensuite à l'entreprise de justifier ses choix en démontrant, étude d'impact à l'appui, que les mesures mises en œuvre étaient raisonnables, même si elles n'ont pas permis de réduire le risque à zéro.

Au bout du chemin, l'application d'un régime strict de responsabilité aux responsables de traitement

Le RGPD impose un régime strict de responsabilité aux responsables de traitement : dès lors qu'une violation est constatée, sa réparation sera automatique. Les personnes concernées peuvent intenter une action sans avoir à prouver une faute ou une négligence de la part du responsable de traitement. La charge de la preuve que « le fait qui a provoqué le dommage ne lui est nullement imputable » [(2)] (à savoir que le traitement de données à caractère personnel est réalisé conformément au RGPD et aux droits nationaux transposant le RGPD) pèse sur le responsable de traitement défendeur. Les sociétés doivent en outre être préparées à ce que les personnes concernées exercent leur droit d'introduire une réclamation auprès d'une autorité de contrôle pour avoir accès aux conclusions de l'enquête administrative. Il est probable que les personnes concernées utiliseront ces informations dans le cadre de procédures civiles. Du fait de cette approche, les personnes concernées peuvent facilement créer une présomption de violation de la protection des données à caractère personnel et une charge administrative encore plus lourde pèse sur les responsables de traitement.

Les dispositions de responsabilisation du RGPD exigent des défendeurs de prouver qu'ils ont mis en œuvre les « mesures techniques et organisationnelles appropriées ». Les responsables de traitement doivent considérer la logique de responsabilisation du RGPD comme une stratégie précontentieuse, conçue pour créer des documents permettant de démontrer que le défendeur a appliqué les mesures techniques et organisationnelles appropriées. Le registre des opérations de traitement et l'analyse d'impact seront déterminants pour renverser la présomption de faute de la part du responsable du traitement.

En conclusion, à l'avenir, on peut s'attendre à ce qu'il y ait une convergence dans les démarches des entreprises en matière de risques RGPD et en matière de risques liés à la sécurité des produits ou à la pollution de l'environnement. Les analyses d'impacts et analyses de risques seront similaires, encourageant une mutualisation d'expertises au sein du groupe sur la gestion des risques et la préparation d'analyses d'impacts. Le programme RGPD de l'entreprise devra naturellement s'intégrer dans le dispositif général de gestion des risques du groupe.

(2) Article 82-3, RGPD.

Données et règles de concurrence

Par Anne PERROT
Associée, MAPP

Les entreprises dont le modèle d'affaires repose sur les technologies numériques sont présentes dans tous les domaines d'activité : transport, hôtellerie et hébergement, banque, culture… Les plateformes numériques ne constituent pas un « secteur » mais, selon les cas, renouvellent une offre existante par les technologies numériques (comme le modèle des VTC concurrençant l'ancien secteur des taxis) ou créent de nouveaux services que seules ces technologies peuvent offrir (comme le guidage des automobilistes en temps réel dans la circulation). Ces entreprises opérant à partir de technologies numériques se trouvent donc dans des positions concurrentielles très différentes selon qu'elles sont de nouveaux entrants dans des marchés déjà constitués ou bien des innovateurs sur des marchés émergents. Pourtant, elles partagent bien un point commun : leurs services sont rendus à l'aide d'algorithmes, dont la matière première est constituée par les données des utilisateurs.

Le rôle des données dans l'offre de certains services n'est pas nouveau. Les statistiques et les données existent depuis longtemps et les entreprises qui les utilisent ne sont pas, elles non plus, arrivées ces dernières années sur le marché : les services de météorologie ou les études de marché reposent ainsi sur l'exploitation de données. La rupture vient plutôt aujourd'hui du fait que ce sont les données des utilisateurs du service eux-mêmes qui sont les ingrédients du fonctionnement de celui-ci. Contrairement à un service météorologique qui utilise des données scientifiques pour fournir ses prévisions, un service de guidage des automobilistes exploite les données de déplacements de ses propres utilisateurs ; un service de comparaisons de produits utilise également les notations des acheteurs passés utilisant le service pour fournir son propre classement ; un moteur de recherche présente comme pertinents les résultats les plus consultés par les internautes en réponse à la même requête.

Contrairement aux services qui utilisent des données statistiques ou scientifiques, ceux qui font usage des données des utilisateurs sont d'une qualité d'autant meilleure que ceux-ci sont nombreux. Il s'agit d'un « effet de réseau direct », par lequel le nombre des consommateurs d'un produit ou d'un service accroît pour chacun d'eux l'utilité retirée de celui-ci, cet accroissement passant ici par l'amélioration de la qualité du service en question. D'autres plateformes fonctionnent sur le principe des « effets de réseaux indirects » : un plus grand nombre de personnes raccordées à « l'autre face » de la plateforme (chauffeurs, hôtels) accroît l'utilité des utilisateurs d'une face donnée (personnes cherchant à se déplacer ou à être hébergées) et *vice versa*. Dans un cas comme dans l'autre, l'une des clés de la réussite de ces services repose sur leur capacité à recruter de nombreux utilisateurs : l'accroissement de la qualité de service (et donc, sur ces plateformes souvent gratuites pour les consommateurs, l'attractivité) repose donc pour les entreprises du numérique sur la grande taille [(1)]. En particulier, l'efficacité des plateformes repose souvent sur la détention de données individuelles collectées en grand nombre (*Big Data*) : qu'il s'agisse de la pertinence des réponses à une requête, de la faculté de proposer des hébergements variés, ou de guider au mieux les automobilistes dans les embouteillages, les algorithmes développés par les plateformes numériques sont d'autant plus performants qu'ils optimisent leurs résultats à partir d'un grand nombre d'utilisateurs.

(1) Voir ROCHET J.-Ch. et TIROLE J., « Platform Competition in two-sided Markets », *Journal of the European Economic Association*, June 2003 1(4):990-1029 ou encore ARMSTRONG M., "Competition in Two-Sided Markets", *The RAND Journal of Economics*, Vol. 37, n° 3 (Autumn, 2006), pp. 668-691.

Ce phénomène est à l'origine d'un grand nombre des problèmes de concurrence posés par ces nouveaux modèles d'affaires. En particulier, l'une des questions est d'évaluer le danger potentiel pour la concurrence que représente la détention de données, danger qui pourrait alors avoir plusieurs sources. Certaines questions sont en effet posées avec une dimension nouvelle par l'économie numérique [2]. La détention de données personnelles sur une plateforme est-elle un frein au changement d'opérateur pour les consommateurs et donc à la liberté de ceux-ci de mettre en concurrence les différents offreurs ? Les données sont-elles susceptibles de constituer une barrière à l'entrée pour les entrants potentiels et sont-elles à la source d'une position dominante pour l'opérateur qui les détient ? Les règles de concurrence actuelles sont-elles capables de traiter les problèmes complexes liés aux données et aux algorithmes ?

Le côté demande : les données et les barrières à la mobilité

Les consommateurs qui rejoignent une plateforme sont souvent « attachés » à elle par plusieurs sortes de liens. Sur les plateformes de vente, il n'est pas rare qu'ils aient la possibilité de rentrer leurs coordonnées de cartes bancaires, ce qui simplifie les achats suivants, et permet aisément, en particulier, de procéder à des achats à partir d'un smartphone. Sur les réseaux sociaux se trouve l'ensemble des relations des internautes et l'historique de leurs échanges, les photos, vidéos, documents partagés, etc. Sur les plateformes permettant des déplacements, les consommateurs peuvent rentrer leurs destinations favorites et accélérer ainsi leurs recherches. Les smartphones ont également de nombreux contenus (musique et vidéos notamment) achetés au fil du temps par les utilisateurs. La conséquence du fait que les consommateurs détiennent des données sur ces différentes plateformes est de renchérir les coûts de changement d'opérateur [3] (*switching costs*) : en effet, reconstituer cet ensemble de données auprès d'une nouvelle plateforme engendre des coûts qui peuvent rendre les consommateurs réticents à changer de plateforme et freiner la mobilité entre plateformes. De ce fait, les secteurs où les consommateurs subissent d'importants *switching costs* sont plus faiblement concurrentiels et permettent aux opérateurs d'extraire des rentes de leurs utilisateurs.

Évidemment, ces différents types de coûts ne sont pas équivalents : si le fait d'entrer des données de carte bancaire ou de destinations sur un site ne représente qu'un coût en temps, qui plus est de faible ampleur, l'achat de morceaux de musique cumulé sur plusieurs années peut représenter un montant monétaire important. De même, l'historique des échanges avec ses relations sur un réseau social et les données personnelles ainsi amassées est impossible à reconstituer et le risque de sa perte représente alors un frein important au changement de réseau.

Ce problème des freins à la mobilité n'est pas nouveau. Ainsi la concurrence bancaire est-elle également limitée par des freins à la mobilité, et l'identification de ce problème a donné lieu à plusieurs mesures de politique publique correctrice (loi Hamon et loi Macron), comme l'obligation faite aux banques (de départ et d'arrivée) de faciliter à leurs clients la mobilité bancaire. L'équivalent dans le secteur numérique est constitué par l'ensemble des mesures permettant la « portabilité » des données, sur le modèle de la portabilité des numéros de téléphone destinée à faciliter le changement d'opérateur et à fluidifier la concurrence. L'article 12 de la loi pour une République numérique, dite loi Lemaire, a ainsi instauré la possibilité pour les consommateurs d'emmener avec eux les données accumulées sur une plateforme [4]. Cette disposition, qui n'a pas pour objectif premier de fluidifier la concurrence mais de permettre aux internautes de maîtriser leurs contenus

(2) Voir STUCKE M. et GRUNES A., *Big Data and Competition Policy*, Oxford University Press, Oxford, 2016.
(3) Voir NASSE Ph. : *Rapport sur les « coûts de sortie »*, 22 septembre 2005, Rapport pour le compte du ministre de l'Industrie.
(4) Cette possibilité sera effective à partir du 25 mai 2018, date d'entrée en vigueur du Règlement européen 2016/679 sur la Protection des Données, qui prévoit la même disposition au plan européen.

numériques, a pourtant une incidence importante en matière de concurrence puisqu'elle réduit les freins à la mobilité et permet aux consommateurs de choisir leur plateforme dans le cadre d'une concurrence « par les mérites ».

Le côté offre : les données comme source de comportements anticoncurrentiels

L'autre question que pose l'accumulation des données par les plateformes numériques est celle des comportements anticoncurrentiels auxquels la détention de données peut spécifiquement donner naissance. Le rapport commun publié en 2016 par l'autorité allemande de la concurrence (le Bundeskartellamt) et l'Autorité de la concurrence française [5] passe en revue les différents mécanismes par lesquels la détention de données peut accroître le pouvoir de marché des plateformes et créer les conditions d'un abus de leur position dominante par les opérateurs.

Les plateformes tendent de fait à vouloir accroître le volume de données dont elles disposent car leur performance, on l'a dit, en dépend directement. Par exemple, la qualité d'un moteur de recherche dépend de la pertinence des résultats qu'il offre aux internautes en réponse à une requête de leur part. La pertinence est améliorée grâce à la « multitude » des utilisateurs auteurs de la requête en question, car c'est leur comportement vis-à-vis des résultats proposés qui aide l'algorithme à améliorer la présentation des résultats. Cette tendance à la grande taille se traduit par l'accroissement du nombre des opérations de concentrations impliquant le secteur des données, passé de 55 en 2008 à 164 en 2012 au sein des pays de l'OCDE. Mais la grande taille, accompagnée éventuellement d'une position dominante, n'est pas nécessairement un facteur d'abus de cette position.

Pour que les données permettent à un opérateur du numérique de s'affranchir de la pression des concurrents, et le cas échéant d'abuser de sa position dominante, il faut que la détention de ces données conduise à l'exclusion de concurrents. Ceci pourrait se produire en cas de refus d'accès à des données indispensables pour l'offre d'un autre service. Mais une telle circonstance ne peut se produire que lorsque les données en question constituent une « facilité essentielle », ce qui est rarement le cas. Les algorithmes peuvent aussi mettre en œuvre une politique de discrimination tarifaire entre les utilisateurs, en s'appuyant sur la connaissance fine de leur demande individuelle pour pratiquer des prix différenciés. Mais là encore, il peut se faire que la discrimination soit en réalité favorable à l'intensité concurrentielle et *in fine* aux consommateurs, en permettant par exemple de proposer des prix plus faibles à ceux qui présentent une faible disponibilité à payer pour le service. Ceci doit donc s'apprécier au cas par cas. Quoi qu'il en soit, s'il est possible d'exclure des concurrents grâce à une tarification discriminante, cette possibilité ne survient que dans des circonstances particulières qui nécessitent un examen au cas par cas.

D'autres facteurs relativisent ce danger d'abus de position dominante. Les effets de réseau peuvent aussi stimuler la concurrence au lieu de l'entraver. Un nouvel entrant proposant un service innovant peut ainsi faire jouer à son profit les effets de réseaux en attirant de nombreux utilisateurs, grâce à la « viralité ». La faculté pour les internautes d'utiliser simultanément les services de plusieurs plateformes (*multihoming* [6]) leur permet de mettre directement différents services en concurrence. Ensuite, les technologies numériques encouragent les innovations, ce qui se traduit par une dynamique concurrentielle importante. Ce flux continu d'entrées d'entreprises utilisant des technologies numériques suggère qu'aucune n'a réellement besoin des données détenues par

(5) *Competition Law and Data*, Rapport commun du Bundeskartellamt et de l'Autorité de la concurrence, mai 2016, http://www.autoritedelaconcurrence.fr/doc/reportcompetitionlawanddatafinal.pdf
(6) Voir par exemple GABSZEWICZ J. J. et WAUTHY X. Y., "Two-Sided Markets and Price Competition with Multi-homing", *Working paper Core*, 2005.

les entreprises en place pour arriver sur le marché. La plupart des études soulignent d'ailleurs que les données nécessaires à une activité (préférences et centres d'intérêt des consommateurs, données de géolocalisation…) peuvent assez facilement être recueillies en grande quantité assez tôt après le lancement d'une offre de service. Par ailleurs, les données utiles à un service peuvent être obtenues de sources différentes. Le rapport du Bundeskartellamt et de l'Autorité de la concurrence cite l'exemple des goûts musicaux des internautes, qui peuvent être appréhendés aussi bien directement par leurs achats sur des sites marchands de musique qu'à partir de la navigation sur des sites de *streaming* ou sur des moteurs de recherche généralistes, ou encore à partir de leurs pages personnelles sur les réseaux sociaux. De tels exemples interdisent de considérer ces données comme relevant d'une facilité essentielle.

Faut-il changer les règles de concurrence pour traiter des entreprises du numérique ?

Certains arguments sont parfois avancés pour préconiser une régulation spécifique aux plateformes ou à tout le moins une adaptation des règles de concurrence aux particularités du numérique. La taille des plateformes, le caractère immatériel de leur activité, l'existence d'effets de réseaux tendant à la constitution de positions dominantes, la détention de volumes considérables de données, tous ces facteurs inciteraient pour les uns à encadrer les pratiques économiques des plateformes par une régulation *ex ante*, pour les autres à modifier les règles de concurrence.

En l'état actuel des savoirs, ces arguments ne résistent pas à l'analyse. Tout d'abord, les comportements des plateformes relèvent bien d'une analyse concurrentielle. Souvent ces comportements se manifestent sur des marchés où plusieurs acteurs sont en concurrence, comme celui des comparateurs de prix, où les opérateurs se battent pour attirer du trafic, ou celui de la publicité, notamment la publicité ciblée, qui voit lui aussi les plateformes se concurrencer. On ne voit pas pourquoi ces marchés, sur lesquels plusieurs offreurs sont en concurrence et mettent en œuvre des stratégies (tarifications particulières, ventes groupées, promotions, etc.) relevant de l'analyse traditionnelle, devraient être soustraits à l'analyse concurrentielle *ex post*.

Le fait que la structure et la technologie des activités numériques s'appuient sur des innovations majeures ne suffit pas à justifier de nouveaux outils. Les phases de transition vers de nouvelles technologies ou de nouveaux modèles d'affaires sont légion et l'analyse concurrentielle sait en général s'en accommoder : ainsi, l'Autorité de la concurrence a, en France, au terme d'une analyse mettant en œuvre de nouvelles méthodes, conclu à l'existence d'un seul marché pertinent incluant ventes en dur et ventes en ligne dans l'opération de rapprochement entre la FNAC et Darty. Les autorités de concurrence ont déjà été amenées à examiner le risque que fait peser sur le fonctionnement concurrentiel des marchés le rapprochement de deux entreprises détentrices de données. L'une des caractéristiques des bases de données est en effet que leur croisement est source d'externalités, la valeur potentielle des données résultant du croisement étant démultipliée par rapport à la valeur des données prises séparément.

La Commission européenne, de son côté, a sanctionné Google pour avoir avantagé, au détriment des autres comparateurs de prix, les services de son propre comparateur Google Shopping.

On peut objecter que le temps d'instruction nécessaire pour parvenir à une décision est trop long par rapport à celui des affaires en termes de numérique. Mais ceci plaide plutôt pour une mise à niveau des compétences des autorités en matière de numérique et notamment pour le recrutement de *data scientists* et d'informaticiens [7].

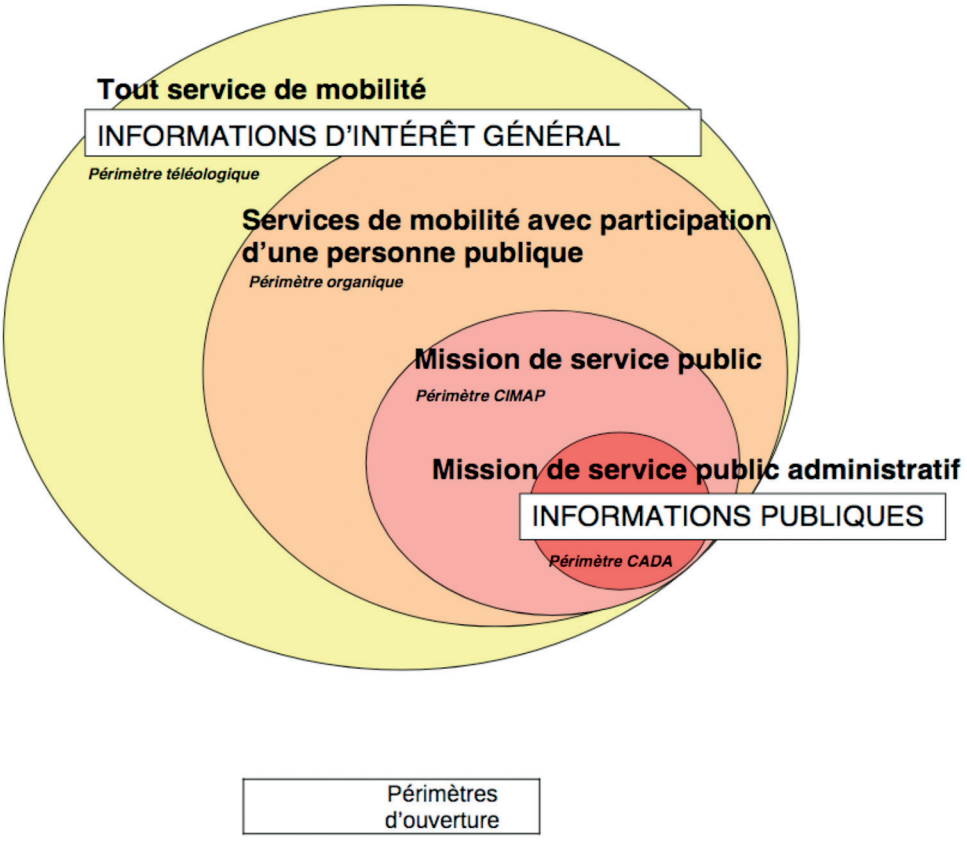
(7) Voir les notes du Conseil d'Analyse économique, n° 26, *Économie numérique* (par LANDIER A., COLIN N., MOHNEN P. et PERROT A.) et n° 36, *Régulation économique : quels secteurs réguler et comment ?* (par BACACHE M. et PERROT A.).

Comment définir et réguler les « données d'intérêt général » ?

Par Bertrand PAILHÈS

En 2015, les données de transport font l'objet d'un débat intense entre les entreprises de transport public (RATP, SNCF, Transdev, etc.), les nouveaux acteurs venus du numérique (Google, Uber mais aussi Citymapper) et les pouvoirs publics, qui souhaitent alors développer l'ouverture des données publiques.

En mars 2015, Francis Jutand remet son rapport sur l'ouverture des données de transport [1], ouvrant la voie à des dispositions législatives qui seront intégrées dans la loi Macron promulguée en août de la même année. Il y définit les « informations d'intérêt général », qui rassemblent toutes « données (ou classes de données) des services de mobilité dont l'ouverture serait jugée opportune ». Y sont incluses des données de services publics mais également d'entreprises privées fournissant un service de mobilité.



Si les dispositions de la loi se sont finalement concentrées sur les acteurs délégataires d'une mission de service public, ce rapport fut le premier, en France, à élargir le débat de l'ouverture des données

(1) https://cdn.nextinpact.com/medias/rapportjutand.pdf

publiques (l'*open data*), en promouvant une vision complète des enjeux autour des données, dans un domaine où l'organisation des acteurs publics et privés est essentielle.

Par la suite, le rapport sur les « données d'intérêt général (DIG) » rendu en septembre 2015 [2] a étendu le cadre d'analyse juridique et économique à tous les domaines en identifiant les leviers et obstacles permettant l'accès à ces données. La communication de la Commission européenne sur l'économie de la donnée publiée en janvier 2017 [3] soulève également la question de l'accès aux données entre acteurs publics et privés. Enfin, la question du partage des données publiques et privées a figuré au centre du rapport Villani et de la stratégie française sur l'intelligence artificielle [4].

Les DIG, pour quoi faire ?

L'exemple des transports est le meilleur exemple illustrant la manière dont les données doivent être collectées et partagées entre acteurs publics et privés. En effet, dans ce secteur :

- les données sont abondantes ;
- l'utilité de leur collecte massive est directement compréhensible pour traiter des problèmes quotidiens que sont les embouteillages, la pollution de l'air ou le fonctionnement des transports publics ;
- elles sont collectées par une grande variété d'acteurs, du plus privé (la voiture individuelle) au plus public (l'infrastructure en site propre sur domaine public, comme le métro), chaque acteur de la chaîne étant relié à ces deux dimensions ;
- le secteur est l'objet de l'attention de nouveaux acteurs perçus comme puissants, notamment les grandes plateformes numériques qui ont compris que la gestion du transport quotidien est une application plébiscitée par les utilisateurs, qu'ils soient en voiture (Waze) ou en transport en commun (Citymapper, Google Maps).

L'enjeu de l'accès aux données est donc double : d'une part développer une activité économique en soutenant l'innovation et en évitant la constitution de rentes et, d'autre part, améliorer la conduite des politiques publiques.

Du point de vue économique, la donnée est en effet un bien non rival, dont l'usage par une personne n'en prive pas les autres : le coût marginal de reproduction étant quasi nul, il est possible, à partir d'une même donnée, d'en multiplier les usages et les utilisateurs et, partant, de maximiser la valeur potentielle qu'il est possible de tirer de cette donnée. Encourager le partage de données entre acteurs est donc d'abord un moyen de développer l'innovation et de garantir une exploitation économiquement efficace de cette ressource que forment les données.

Par ailleurs, la maîtrise des données importantes pour l'économie et la société est un enjeu croissant identifié par les pouvoirs publics pour constituer les entreprises et services leaders de demain, dans l'ensemble des secteurs économiques. Ainsi, vouloir développer une industrie européenne de l'intelligence artificielle [5], qui repose nécessairement sur de larges quantités de données, nécessite de traiter la question de l'accès aux données détenues par les entreprises ou les administrations quand celles-ci en ont le monopole [6].

Du point de vue politique, le débat sur les données d'intérêt général rejoint la vision du numérique comme « commun », c'est-à-dire comme un ensemble de ressources n'appartenant à personne et

(2) https://cdn2.nextinpact.com/medias/rapport-cytermann.pdf
(3) Communication de la Commission européenne du 10 janvier 2017 « Créer une économie européenne fondée sur les données », http://europa.eu/rapid/press-release_IP-17-5_fr.htm
(4) https://www.aiforhumanity.fr/
(5) Voir notamment le rapport Villani.
(6) Par exemple, dans le cas des entreprises privées, les données de navigation Internet détenues par Google ou Facebook et, dans le cas des administrations françaises, les données de santé.

utilisable par tous, à condition d'en fixer des règles de gouvernance partagées : le logiciel libre, les protocoles Internet ou Wikipédia en sont les exemples les plus emblématiques. La donnée constitue un des enjeux les plus importants des prochaines années pour porter cette vision, qui se rapproche de celle portée par la communauté scientifique pour le savoir scientifique. L'objectif est d'accroître la transparence et ainsi le contrôle des entités détentrices des données pour, *in fine*, garantir la confiance des utilisateurs dans les outils numériques.

Les DIG, quelle définition ?

Les données d'intérêt général (DIG) sont donc des données, publiques ou privées, dont le partage et l'ouverture sont « d'intérêt général », c'est-à-dire d'un intérêt plus large que le seul intérêt du détenteur des données.

Il est essentiel de distinguer cette notion de celle de « l'*open data* » où le législateur poursuit un double objectif de transparence et d'innovation en ouvrant l'accès aux données publiques et en limitant au maximum les conditions de réutilisation (licence, prix, etc.), sans vérifier *a priori* la qualité de la réutilisation. L'*open data* trouve ainsi sa source dans le droit constitutionnel du citoyen à connaître de l'administration ; ce n'est pas le cas des DIG, dont l'accessibilité doit pouvoir être soumise à des conditions particulières et ne constitue pas un droit fondamental.

Il convient également de distinguer les DIG de l'accès administratif aux informations détenues par les personnes publiques ou privées : pour collecter l'impôt, l'administration doit accéder à certaines informations, tout en en garantissant la confidentialité. Les données constituent un nouveau champ d'application de cette prérogative de puissance publique.

Dans les deux cas précédents, une raison impérative, d'intérêt général et imposée par le législateur, définit le régime d'accès aux données, et il ne semble pas nécessaire de développer de nouveaux concepts pour mettre en œuvre cet accès.

Les DIG, à l'inverse, peuvent constituer un outil pour faciliter le partage de données dans des cas plus complexes, où la prise en compte des spécificités du cas est nécessaire, sous réserve de lever les obstacles techniques, juridiques ou économiques.

Les DIG, quels obstacles ?

Obliger une personne détenant des données à les rendre accessibles à des tiers suppose d'abord de tenir compte des contraintes juridiques qui peuvent empêcher cet accès.

Contrairement à une idée répandue, il n'existe pas aujourd'hui de droits de propriété sur les données détenues par une personne : des droits de propriété intellectuelle s'appliquent aux bases de données (droit *sui generis*) mais ils ne s'étendent pas aux données elles-mêmes. Pour autant, le détenteur des données peut faire valoir des secrets légaux comme la protection des données personnelles, ou des obligations comme la sécurité informatique pour justifier une limitation d'accès aux données. Ces considérations, légitimes, doivent être prises en compte dans les modalités d'accès aux données, au travers par exemple des API [7] et des procédures de sécurité comme la traçabilité des accès.

De telles règles d'ouverture pourraient également limiter la liberté d'entreprendre de l'entreprise détentrice. Si ce principe constitutionnel est important, il est toutefois possible, en France, d'en limiter les effets à condition de justifier d'un motif d'intérêt général qui surpasse la protection de cette liberté. La question de la proportionnalité est ainsi centrale quand il s'agit de définir l'accès aux DIG.

(7) Application Programming Interface : interface standardisée d'accès à des informations ou des fonctionnalités.

Sur un plan économique, malgré le caractère non rival de la donnée, l'obligation de rendre des données accessibles à des tiers peut réduire l'incitation du détenteur de la donnée à investir, dès lors qu'il sait qu'il sera contraint de partager la donnée et n'en aura pas le monopole. C'est pourquoi l'accès aux DIG ne doit pas nécessairement être gratuit mais fixé à un prix qui maintient l'incitation du détenteur à investir, voire constitue même un revenu complémentaire pour celui-ci. Bien évidemment, ce prix peut également dépendre de la nature du demandeur, un laboratoire de recherche publique n'ayant pas la démarche commerciale d'une *start-up*.

L'ouverture des DIG s'entend donc uniquement dans un environnement technico-économique maîtrisé par le détenteur, qui doit être en mesure de spécifier les conditions techniques d'accès aux données (format, interfaces ou API, et mesures de sécurité), mais également les conditions d'utilisation des données ainsi récupérées, leur utilisation étant, de toute façon, soumise aux règles de droit commun applicables (données personnelles, réglementations sectorielles).

Dès lors, comment organiser l'accès aux données d'intérêt général ?

Écartons d'emblée la fausse piste d'un droit de propriété intellectuelle associée à la donnée. Cette idée est séduisante par son analogie avec d'autres biens immatériels (brevets, œuvres) et semble permettre d'envisager une régulation « économique » de l'accès aux DIG : si un « propriétaire » de données a un intérêt économique à les ouvrir en les valorisant, il le fera par le jeu du marché. Mais de même que certaines infrastructures sont plus efficacement construites et entretenues par la puissance publique sans appropriation par un acteur particulier, certaines données, de nature stratégique ou indispensables pour comprendre et résoudre certains problèmes sociaux ou économiques, ne doivent pas être exclusivement détenues par un acteur.

L'exemple de la voiture connectée illustre cette impasse de la propriété : si le constructeur est « propriétaire », assureurs, sociétés d'autoroute, réparateurs, calculateurs de mobilité mettront en place de multiples dispositifs pour collecter directement les données plutôt que de risquer un prix trop élevé de la part du constructeur. Il semble ainsi beaucoup plus pertinent de prévoir que les données de l'automobile seront accessibles à tous ces acteurs, dans des conditions différenciées selon la nature du besoin de chacun.

Se pose ensuite la question de la définition d'un régime général : peut-on construire juridiquement cette notion de DIG sans s'accrocher à des considérations sectorielles (transport, logement, écologie) pour lesquelles il semble plus simple de qualifier à la fois les objectifs poursuivis et les données visées ? Le premier rapport de septembre 2015 sur les DIG invitait les pouvoirs publics à suivre cette voie, hormis dans deux cas : l'accès de la statistique publique aux données des entreprises dans le cadre d'enquêtes [8] et l'extension de l'ouverture des données publiques aux entreprises privées ayant une mission de service public [9]. À la suite de la loi Macron de 2015, la loi pour une République numérique d'octobre 2016 a ainsi poursuivi l'accès à certaines données spécifiques, dans les domaines de l'énergie ou des vitesses maximales autorisées. La loi Santé a également engagé le mouvement de partage et d'ouverture des données de santé avec la création du système national des données de santé.

Il semble toutefois possible de développer un modèle général de la donnée « d'intérêt général », en répondant aux quatre questions suivantes :

- Quelle est la donnée visée, ou le type de donnée, et qui est son détenteur ?
- Qui demande l'accès aux données ?
- Pour quelle finalité une personne souhaite-t-elle accéder aux données ?
- Selon quelles modalités l'accès est-il fourni ?

(8) Introduit dans l'article XX de la loi pour une République numérique.
(9) Déjà prévu dans la loi CADA régissant l'accès aux documents administratifs.

Les cas où il suffit de répondre à une seule de ces questions pour obtenir l'accès à des données sont rares [10] et c'est en combinant ces critères que l'accès d'un tiers à des données pourra être permis. Ainsi, un laboratoire de recherche publique pourrait faire valoir sa qualité pour accéder à des données détenues par une entreprise mais il devra justifier *a minima* de modalités sécurisées d'accès. De même, des données de transport d'une compagnie de taxis pourraient être *a priori* considérées comme d'intérêt général mais l'étendue et les conditions d'accès pourraient en être différenciées en fonction de la nature du demandeur et de la finalité qu'il poursuit : l'accès de la Mairie de Paris n'est pas le même que celui d'un service de calculateur d'itinéraires.

Pour traiter cette complexité casuistique, un régime général pourrait définir deux principes de régulation :

- la nécessité de faire prévaloir le contrat entre les parties sur toute intervention publique ;
- l'obligation de faire droit aux demandes d'accès raisonnables pour les détenteurs de DIG [11].

Le premier principe est évident quand on constate que le partage des données n'est pas une idée nouvelle dont les pouvoirs publics auraient le monopole : de nombreuses entreprises ont l'idée de mettre en commun leurs données ou d'en organiser le partage afin de permettre à des tiers d'imaginer de nouvelles applications ou de conduire des recherches. Récemment, Transdev a lancé une plateforme, « Catalogue », dont l'objectif était de rassembler les données de tous les opérateurs de transport, de BlaBlaCar à la SNCF. Uber a lancé « Uber Movement » pour permettre à des tiers d'accéder aux données de ses taxis.

Ces exemples montrent la prise de conscience des acteurs économiques, qui utilisent également les données pour construire un écosystème d'autres acteurs économiques dépendant d'eux. Mais ils montrent aussi le besoin de régulation : l'initiative de Transdev n'a pas réussi à convaincre la SNCF et la RATP de fournir leurs données et Uber est aujourd'hui seul à choisir quelles sont les données qu'il ouvre.

La puissance publique a donc un premier rôle d'arbitre à endosser, pour permettre l'accès de tiers, « dans des conditions raisonnables », qui peuvent être techniques, juridiques et économiques. Ces conditions tiendraient également compte des quatre questions mentionnées ci-dessus et notamment de la nature du demandeur et de la finalité poursuivie.

Afin de renforcer les possibilités d'accès, notamment aux données des acteurs privés, les pouvoirs publics devront également établir *a priori* les critères de définition des DIG (en les déclinant éventuellement par secteur). Parmi ces critères généraux, on peut évoquer l'importance des données pour un enjeu de politique publique (sécurité, santé, environnement, etc.), le caractère unique ou universel du jeu de données (notamment pour le développement économique) ou encore la contribution des utilisateurs à la constitution du jeu de données (dans le cas de Waze par exemple). Sur la base de ces critères, l'autorité publique engagerait une procédure transparente et contradictoire visant à déclarer «d'intérêt général » un ensemble de données et ouvrant ainsi droit aux demandes d'accès.

Ces pistes, déjà explorées dans les travaux cités en introduction de cet article, gagneraient enfin à être harmonisées au niveau européen afin d'éviter tout risque d'évasion réglementaire des acteurs concernés.

Le besoin de régulation de l'accès aux données publiques et privées « d'intérêt général » est de plus en plus important et il semble donc possible de définir un cadre simple, flexible et générique qui préserve les intérêts des parties tout en garantissant une circulation et une ouverture accrue des données.

(10) C'est le cas de la statistique publique (finalité qui permet un accès généralisé), de l'autorité judiciaire (qui a un pouvoir général d'enquête) ou de l'ouverture des données publiques (la nature du détenteur permet l'accès inconditionnel).

(11) Cette proposition est directement inspirée de l'obligation d'accès imposée à l'opérateur historique dans le secteur des télécommunications.

Éthique et Big Data : désenchanter le numérique

Par Jean-Baptiste SOUFRON
Avocat associé
FWPA Avocats

L'entrée des individus dans la société des données s'est accomplie autour de deux malentendus importants. Le premier est que la révolution numérique serait d'abord une révolution technologique et sociale, et le second que les technologies sont neutres et ne sont que des outils dont les conséquences dépendent entièrement de la volonté de leurs utilisateurs.

Or, l'objectif des technologies numériques n'est pas de changer la société.

Il suffit pour s'en convaincre de se rappeler que l'article séminal de Vannevar Bush en 1944 était intitulé "How we may think". L'ouvrage fondateur de Norbert Wiener en 1950 s'intitulait *The Human Use of Human Beings*. C'est auprès de l'« Augmentation Research Center » que Douglas Engelbart a présenté, en 1968, ses travaux les plus importants.

C'est bien l'homme qu'il s'agit de transformer, en lui attribuant de nouvelles capacités par le biais de machines de plus en plus légères et de plus en plus connectées. De ce point de vue, la révolution sociale du numérique n'est qu'une conséquence de la tentative de création d'un *homo numericus* – un être intégré dans un ensemble de boucles cybernétiques au sein desquelles les choix individuels sont limités, rationalisés, contrôlés.

Quelle que soit la réalité – limitée pour dire le moins – des accomplissements de ce projet transhumaniste, pour ne pas dire eugéniste, il constitue le nœud du problème éthique au cœur du numérique.

Car l'avènement de l'homme augmenté ne relève pas d'un projet simplement philosophique, mais d'un projet industriel et politique au sens le plus concret du terme. D'un côté, appliquant à la lettre les enseignements de la théorie de l'information qui facilite la division de la pensée, il est présenté comme ubiquitaire et vise à toucher l'ensemble de la population. De l'autre, se présentant comme doté d'une rationalité et d'une puissance explicative plus profonde que le reste de la science, il est présenté comme transcendant, et bien sûr omnipotent.

Autrement dit, l'éthique de l'ubérisation ne s'embarrasse pas de détails. Sous couvert de progrès et de rationalité, elle a vocation à s'appliquer à tous, à régir chaque aspect de leur vie, à transformer les règles sociales, voire à favoriser l'émergence d'une morale et d'une spiritualité nouvelles, plus adaptées à ce nouvel environnement.

Ce faisant, elle bouscule à la fois les racines de l'individu et les fondements de la vie collective, notamment l'idée qui remonte à John Locke et selon laquelle la légitimité des règles communes repose nécessairement sur une forme ou une autre de consentement collectif.

Cette analyse se traduit par une remise en cause régulière de la capacité de l'État à accompagner l'individu pour lui permettre d'accomplir son potentiel. Une récente étude de Stanford montre ainsi que les dirigeants de la Silicon Valley s'identifient essentiellement comme des Démocrates poursuivant des objectifs de gauche et souhaitant permettre le développement et la valorisation des individus, mais qu'ils refusent de le faire en adoptant des outils collectifs tels que les syndicats, le solidarisme ou la redistribution fiscale – notamment en matière de droit du travail.

Ce qui revient en définitive à vouloir assurer la justice sociale tout en dérégulant au maximum la société – un paradoxe qui ne fait sens que lorsqu'on comprend que l'objectif du projet est de transformer les individus pour assurer automatiquement la justice sociale grâce à la régulation intelligente de leurs interactions par la donnée.

Aux yeux de ces dirigeants de la Silicon Valley, si les gens ont besoin de l'État, d'une représentation parlementaire, de tribunaux, de corps intermédiaires, de protections et d'encadrement… c'est qu'ils ont besoin d'être améliorés.

Or, de ce point de vue, force est de constater que le projet numérique est un échec. Malgré les ambitions paternalistes des nouveaux barons industriels du numérique, leurs services se révèlent attiser la violence et le mensonge tout autant qu'ils facilitent la communication et le dialogue.

Deux phénomènes se rejoignent.

D'une part, les individus sont désormais contraints de vivre une grande partie de leur vie par l'intermédiaire de leur double numérique, c'est-à-dire par l'intermédiaire d'une hallucination cognitive qui les soumet pleinement aux feux de l'envie – Instagram, Snapchat ou Facebook n'apparaissant alors comme rien d'autre que la mise en application cybernétique des théories du désir mimétique.

D'autre part, et c'est logique, les individus se rendent malades. Si la généralisation des données parvient à réduire les distances et à faciliter la transmission des connaissances, elles n'en transforment pas pour autant les êtres humains. Il faut admettre que notre structure cognitive est en grande partie acquise, et que son accélération ou son augmentation ne se traduisent pas forcément par une amélioration individuelle notable, et encore moins une amélioration sociale. Savoir plus et mieux ne signifie pas penser plus et mieux. Le *nudge* [1] et l'automatisation ne peuvent pas remplacer la morale et le travail.

Pour le dire autrement, le numérique et les données entraînent une standardisation massive et permanente. Ce qui se traduit au niveau économique par la reconstitution des oligopoles de l'information et de la communication correspond, au niveau individuel, à l'incapacité de prendre en compte les exceptions et la créativité. Rien ne ressemble plus à un mail qu'un autre mail, à un flux de réseau social qu'un autre réseau social, à un like qu'un autre like.

Naturellement, ce n'est pas une fatalité. Le paysage numérique a déjà prouvé qu'il pouvait accueillir une part d'humanité. Malgré leur apparence répétitive, le contenu des pages de Wikipédia révèle rapidement à la fois les passions et les faiblesses de ceux qui les écrivent. Les blogs sont un foisonnement curieux et brillant qui recèle une richesse sans limite. De DeviantArt à YouTube, de nombreux sites sont des havres de créativité pour des âmes d'artistes qui n'auraient jamais pu trouver à s'exprimer aussi facilement ailleurs.

Mais le problème n'en est que plus pressant. Et il est encore accentué par l'absence même de reconnaissance de son existence. À entendre les positions des uns et des autres, *There Is No Alternative*. En s'alliant au grand mouvement de la globalisation en en reprenant – au moins de façon superficielle – les théories libertariennes de l'école de Chicago, l'industrie des données a mis en place sa propre version de la *Shock Doctrine* qu'elle administre avec passion à chaque nouvel État qui lui en fait la demande. À chaque problème, sa solution. À chaque solution, ses données. À chaque donnée, son modèle économique.

Le simple fait de remettre en cause le dogme systématique des données semble aujourd'hui relever de l'énormité, car les technologies n'auraient pas d'autres sens que ceux que veulent bien leur donner leurs utilisateurs.

(1) Influence sans contrainte.

Les technologies ne sont pourtant pas neutres, et encore moins quand il s'agit de technologies qui touchent d'aussi près le fonctionnement cognitif humain que les technologies de données.

C'est en grande partie à tort que le mythe fondateur de la Silicon Valley donne l'impression que les outils que nous utilisons aujourd'hui ont été inventés et développés par des acteurs du privé – des entrepreneurs représentant à la fois le rêve américain et l'idéal scientifique d'une nation tournée vers le progrès.

Mais comme l'explique fort bien Alexander Klimburg dans son récent ouvrage *The Darkening Web*, il faut toujours garder à l'esprit l'importance du rôle des militaires dans le développement de l'industrie des données. Il faut avoir conscience de ce que nos vies en ligne ne seraient pas possibles sans la commercialisation des innovations militaires.

C'est bien en effet la *Defense Advanced Research Projects Agency* (DARPA) américaine qui a sponsorisé la recherche et le développement de l'Internet, des interfaces graphiques qui nous permettent d'interagir avec nos appareils, de nombreux outils d'intelligence artificielle et des technologies de reconnaissance vocale, voire des polymères à haute performance indispensables aux écrans de nos téléphones portables.

L'armée est également extrêmement présente au niveau du financement de ces technologies. Avec 2,5 milliards de dollars par an apportés à 97 % par les agences de sécurité, le programme *Small Business Innovation Research* (SBIR) représente la plus importante source de financement pour les entreprises – d'autant plus importante qu'elle sert non seulement de « certification » gouvernementale pour les investisseurs privés, mais aussi d'incitation à l'entrepreneuriat puisque c'est l'un des rares financements qui n'exigent pas d'apport en capital en échange du versement des fonds.

À titre d'exemple, et pour montrer à quel point cette question touche l'ensemble des sociétés de ce secteur, on peut se référer à Mariana Mazzucato qui a remarquablement examiné le cas d'Apple dans son ouvrage *The Entrepreneurial State*. En effet, tout en ayant le plus faible montant de dépenses de recherche et développement des *Big Five* du numérique, la société a réussi commercialement en intégrant des technologies financées par l'armée et par des agences de renseignement, comme les écrans tactiles et la reconnaissance faciale, dans des produits commerciaux élégants et attrayants.

Comme Jacques Ellul ou Alexandre Grothendieck, les premiers à avoir pointé les impasses éthiques du numérique étaient aussi les premiers à comprendre que les technologies ont un sens, et donc qu'elles entraînent une responsabilité.

Incapables de refaire la part des choses entre la morale du numérique et sa réalité, certains individus réagissent désormais de façon violente à des technologies dont le sens leur paraît intolérable. On jette des pierres sur les bus de Google à San Francisco. En 2016, à Nantes, la Cantine, un fablab et incubateur, a été incendiée.

En novembre 2017, à Grenoble, la Casemate, un autre fablab, a été vandalisée et incendiée car décrite comme une institution notoirement nuisible par sa diffusion de la culture numérique. Pire encore, en avril 2018, Nasim Aghdam, une youtubeuse iranienne vivant aux États-Unis, s'est rendue dans les locaux de YouTube pour tirer sur les employés car la société empêchait ses chaînes de recueillir des vues.

Ce n'est pas la première fois que les gens se tournent vers la violence au prétexte de protester contre l'automatisation, le numérique et les données.

On peut se rappeler l'apparition en France, de 1979 à 1983, du Comité de Liquidation ou Subversion des Ordinateurs (CLODO), qui fut actif dans la région de Toulouse, où ses membres posaient des bombes et brûlaient des bâtiments (CII-Honeywell Bull en 1980, International Computers Limited en 1980, Sperry-Univac en 1983, etc.).

À l'époque, ces terroristes numériques expliquaient aux médias français qu'ils étaient des travailleurs dans le domaine de l'informatique, donc bien placés pour connaître les dangers actuels et futurs de l'informatique et des télécommunications. Selon eux, l'ordinateur est l'outil favori des dominants. Il est utilisé pour exploiter, ordonner, contrôler et réprimer.

Mais ils n'étaient pas seuls. En 1983, en Allemagne de l'Ouest, un centre de conception de logiciels utilisés dans les missiles Pershing avait été détruit par un groupe appelé Rote Zellen. En 1984, un groupe belge appelé Communist Combattant Cells (CCC) avait détruit par bombe le siège de plusieurs entreprises en Belgique et en Allemagne. À Londres, un groupe appelé Angry Brigade avait essayé de faire de même. Et il y a eu des actions similaires en Asie, en Amérique du Sud et bien sûr aux États-Unis.

Ces actions ont été souvent qualifiées de violences luddites, mais rien n'est plus faux puisque ce n'étaient pas leurs emplois que les auteurs des violences estimaient menacés par le numérique ou leurs données.

En août 1983, le CLODO a par exemple donné une rare interview en anglais à *Processed World*, où ses membres expliquaient que leurs actions n'étaient ni rétrogrades, ni nouvelles. Selon eux, en regardant le passé, nous ne voyons que l'esclavage et la déshumanisation, à moins de revenir à certaines sociétés dites primitives. Et même si nous ne partageons pas tous le même « projet social », nous savons qu'il est stupide d'essayer de remonter le temps. À leurs yeux, le problème vient plutôt de ce que ces outils sont pervertis à leur origine, ce qui explique par exemple que le secteur le plus informatisé est l'armée, et que 94 % du temps informatique civil est utilisé pour la gestion et la comptabilité.

Pour dire les choses clairement, et pour reprendre les mots des opposants au numérique de 1983, si les microprocesseurs créent du chômage au lieu de réduire le temps de travail de chacun, c'est parce que nous vivons dans une société brutale, et ce n'est en aucun cas une raison pour détruire les microprocesseurs.

C'est avec le film *Wargames* en 1983 que les gens ont dissocié le terrorisme informatique et la violence. Avec l'invention de la figure du hacker – un objet transitionnel permettant de positiver la rébellion au numérique, la politique numérique, les protestations et la violence ne semblent plus se dérouler que dans un monde virtuel et appartenir à une zone grise où les valeurs morales sont distantes et floues. Le choix même des dénominations de *White Hat* – un bon hacker – ou *Black Hat* – un mauvais hacker – semble plus lié au *Seigneur des anneaux* qu'au *Manifeste du parti communiste*.

À cet égard, le livre fondateur de Steven Levy, *L'Éthique des hackers*, marque également un tournant. D'abord par son titre anglais original *Hackers: Heroes of the Computer Revolution*, dont la traduction française montre bien la manière dont la réponse apportée à ces importantes questions se révèle à la fois superficielle et définitive : l'éthique des données ? C'est l'éthique des héros ! Sabre au clair, baïonnette au canon.

Il n'est dès lors pas étonnant que ces analyses soient dénoncées comme un mensonge flagrant par le collectif de Grenoble, ce qui ne justifie évidemment pas le recours à la violence.

Que reste-t-il finalement de révolutionnaire ou de prophétique dans une industrie qui repose sur le capitalisme à l'ancienne (si Uber était vraiment innovant, la société appartiendrait à ses salariés plutôt qu'à ses actionnaires), les monopoles, les micro-tâches, l'argent, etc. ? Quant aux vrais héros, il faut se souvenir que la responsabilité du MIT – pourtant supposé être un parangon d'innovation – a été mise en cause dans le suicide du jeune Aaron Swartz, maltraité pour avoir essayé de hacker une base de données d'articles scientifiques.

Un autre numérique était bien sûr possible. Et il l'est encore. Il ne faut pas oublier que le World Wide Web a été inventé par le CERN entre 1987 et 1993 et que c'est avec lui que démarre une grande vague d'innovation – bien plus qu'avec Internet en tant que tel.

La première étape d'une éthique des données semble finalement de commencer par considérer que les outils numériques ont un sens.

Il faut ensuite avouer que le projet de transformation de l'individu n'a pas de sens, qu'il s'agisse de sa vision la plus violente et eugéniste, ou qu'il s'agisse simplement du *nudge*, son avatar technocratique.

En réalité, et cela ne devrait en fait surprendre personne, le numérique exige une grande rigueur, tant intellectuelle que morale. Le numérique, oui, les données, d'accord, mais pour quoi faire ? Et avec qui ? À défaut, le risque est de se retrouver dans un univers où le choix des données à exploiter ainsi que des algorithmes pour les traiter révèlent les manipulations, ou au moins les biais sous-jacents, de leurs concepteurs.

Malheureusement, pour ce qui concerne la France et l'Europe, la technologie, les outils numériques et l'économie des données semblent surtout avoir été utilisés pour déréguler l'industrie et les services publics. Les postes et les télécommunications qui étaient gérées par l'administration sont désormais largement remplacées par le courriel et la messagerie qui relèvent d'entreprises privées. Les sociétés de presse et d'audiovisuel qui opéraient dans un espace réglementé par des textes aussi importants que la loi de 1881 sur la liberté de la presse ou celle de 1986 sur la liberté de communication sont également concurrencées par des entreprises privées qui n'obéissent qu'aux règles très légères prévues par la loi pour la Confiance dans l'Économie numérique (LCEN, 2004).

Au-delà de cet horizon, on ne perçoit pas en France de vision industrielle claire du numérique et des données.

Aujourd'hui, avec le scandale Cambridge Analytica, même un pilier aussi important de notre société que le processus électoral se retrouve dérégulé par la technologie et les données. Très clairement, le risque est de voir le capitalisme numérique prendre racine dans un anarcho-capitalisme qui efface les frontières, soumet les États et démantèle les règles protectrices des trois marchandises fictives identifiées par Karl Polanyi : la nature, le travail et la monnaie.

Heureusement, les grands axes de réflexion, puissants, remontent déjà à plus de quarante ans puisqu'ils ont été traduits par la loi Informatique et Libertés en 1978. Ils ont démontré leur force dans la mesure où le récent Règlement général sur la Protection des Données européen n'en est que la continuation, de grandes sociétés étrangères comme Facebook ayant d'ores et déjà annoncé qu'elles l'appliqueraient globalement pour le monde entier.

Les principes sont simples. Les données sont qualifiées de personnelles parce qu'elles permettent à l'individu de se définir en tant que personne dans un environnement virtuel. En les protégeant, c'est donc directement la personne des citoyens que l'on protège. Et pour ce faire, c'est le contrôle qui sert de clé. Le consentement est obligatoire. Il est toujours possible d'avoir accès à ses données, de les modifier, ou de s'opposer à leur traitement.

Par ailleurs, ce qui est remarquable pour un texte datant de plus de quarante ans, les décisions uniquement automatisées sont interdites par principe en matière judiciaire, mais également administrative ou privée – pour autant que ces décisions aient des conséquences juridiques pour les individus. C'est d'ailleurs sur la base de ce texte qu'a été censuré en 2017 l'algorithme APB qui servait à orienter les bacheliers dans leur sélection universitaire – comme quoi le droit n'a pas tant de mal à suivre les évolutions de l'innovation.

Autrement dit, si l'éthique du Far West de la Silicon Valley a permis aux États-Unis de développer une industrie des données, c'est l'éthique sociale, solidariste et personnaliste européenne qui peut réussir à la réguler.

À cet égard, les récentes propositions du rapport Villani arrivent à point nommé : audit des intelligences artificielles, évaluation citoyenne, travaux de recherche sur l'explicabilité, éducation et formation à l'éthique dans les écoles d'ingénieurs, étude d'impact, droits collectifs, actions de groupe, observatoire sur la non-prolifération des armes autonomes, comité éthique, etc.

Cependant, elles apportent un certain nombre de solutions qui peuvent être opérationnelles d'un point de vue de politique publique, mais qui ne répondent pas aux questions posées par l'accroissement de l'entropie individuelle générée par le passage dans une société de données, par le droit de pouvoir rejeter certaines technologies en fonction de leur sens et pas seulement sur des critères d'efficacité, par la protection du faible contre le fort, mais aussi du faible contre la foule, etc.

Un premier pas dans cette direction serait la restauration du principe de démocratie – singulièrement mis à mal informellement par les déclarations de mépris ou de prise de contrôle émanant de certains grands dirigeants du numérique, et mis à mal plus formellement par les atteintes directes dont il est l'objet à travers les campagnes de manipulation à grande échelle qui sont désormais orchestrées lors de chaque grande élection occidentale.

Un deuxième pas serait sans doute de retravailler les immenses possibilités encore inexploitées qui sont aujourd'hui court-circuitées par les plateformes et les réseaux sociaux, et qui sont sources d'une grande désillusion.

Naturellement, il faut se pencher sur les enjeux personnels, individuels et familiaux. Il n'est pas normal que la société des données se traduise par une souffrance psychique, par un détachement des liens familiaux, par une pression sociale plus forte et plus violente, voire par une violence tout court. Il faut apprendre à gérer la question du double numérique en permettant qu'il ne soit pas nécessaire de produire constamment des données ou de se soumettre à leur traitement pour pouvoir vivre en société.

Enfin, il faut s'intéresser à la question du sens. Ce qui revient probablement à revenir à des outils fondamentaux tels que le fait d'autoriser ou d'interdire un algorithme, exactement comme la loi de 1978 interdisait déjà de collecter les opinions politiques ou les données ethniques. Toutes les données sont-elles bonnes à prendre ? On peut en douter.

Les données au cœur de la lutte contre la délinquance

Par Éric FREYSSINET
Colonel

Dès la fin du XIXe siècle, la collecte de données s'est imposée avec Alphonse Bertillon comme une des clés de la réussite des enquêtes judiciaires. Au XXIe siècle, la collecte, l'analyse et la présentation des données comme preuves au procès pénal sont au cœur de la lutte contre la délinquance. Elles se concrétisent dans le champ de la criminalistique numérique et du renseignement criminel et leur plein développement passera par une véritable maîtrise des données.

La criminalistique numérique

La criminalistique numérique (ou sciences forensiques numériques) regroupe l'ensemble des techniques de preuve utilisées dans les enquêtes judiciaires et reposant sur des données et supports numériques. Au passage, il est important de constater qu'il existe une forte proximité entre ces méthodes et celles utilisées par les personnes menant des investigations dans le domaine de la sécurité des systèmes d'information.

Évolutions technologiques de la criminalistique numérique

Émergents au début des années 1980, les éléments de preuve numérique sont désormais potentiellement omniprésents, quelle que soit la nature de l'infraction, avec le développement des techniques et des usages. Ils se présentent la plupart du temps sur des supports matériels, mais peuvent aussi être collectés dans l'environnement électromagnétique (comme pour la détection de communications Wi-Fi dans le domicile d'un suspect) ou encore être récupérés chez des tiers (parce qu'hébergés sur des serveurs ou détenus par un opérateur de communications électroniques).

Ces éléments de preuve numériques se caractérisent en outre par leur volatilité et leur potentielle fragilité. Ainsi, le contenu d'une mémoire vive informatique est constamment modifié par le système en fonctionnement, les journaux d'activité d'un serveur informatique font l'objet de rotations et sont régulièrement effacés, quand ce n'est pas la législation qui impose des durées maximales de conservation [1]. Enfin, ils peuvent être fragilisés par la nature des supports de stockage : supports magnétiques et optiques sensibles à leur environnement, ou encore mémoires informatiques susceptibles de générer des erreurs de lecture et d'écriture. Même si ces circonstances sont extrêmement rares et prises en compte par des dispositifs de correction d'erreur, elles doivent être connues de ceux qui interprètent les données numériques dans les enquêtes judiciaires.

Outre les nouveaux supports et les nouveaux environnements, les spécialistes de l'investigation numérique ont dû s'adapter à plusieurs évolutions : la nécessité de collecter des preuves sur des systèmes en fonctionnement (techniques dites de *live forensics*, nécessaires pour les questions de volatilité évoquées plus haut), le développement de l'usage des techniques cryptographiques (et la nécessité qui en découle de mettre en œuvre des méthodes de déchiffrement voire de cryptana-

(1) En France, la durée de conservation des données de connexion par les opérateurs de communication électronique est fixée à un an.

lyse), l'augmentation exponentielle des volumes de données à traiter et la collecte d'éléments de preuve directement sur les réseaux.

C'est bien ce domaine de la collecte de données sur les réseaux et sur Internet qui fait l'objet des développements les plus importants avec l'émergence de nouvelles méthodes et de nombreux outils qui s'avèrent nécessaires pour appréhender les différents territoires qui constituent Internet, tels que les forums et réseaux sociaux, ou encore les réseaux sécurisés ou anonymes autrement appelés *darknets* où les enquêteurs utilisent des techniques d'anonymisation.

Ensuite, l'échange de données avec des acteurs externes aux services d'investigation judiciaire, telles des entreprises de sécurité informatique ou des équipes de réaction aux incidents informatiques (CSIRT), devient de plus en plus courant, ce qui suppose d'utiliser des formats d'échange communs.

Au final, c'est la capacité à traiter, croiser et analyser des volumes de données importants qui devient la plus prégnante. Et donc au-delà de la collecte dans des conditions garantissant l'intégrité des éléments de preuve et l'analyse simple de ces données, ce sont de véritables systèmes de traitement de mégadonnées (*Big Data*) qui sont progressivement mis en œuvre.

Évolutions juridiques et normatives du traitement de la preuve numérique

La législation s'est progressivement adaptée pour accepter les éléments de preuve numériques dans les procédures judiciaires. Ainsi le Code de procédure pénale français (Thiérache et Freyssinet, 2018) prévoit la possibilité de copier des données depuis un support original au moment de la perquisition, de réaliser des enquêtes sous pseudonyme ou encore de collecter des données à distance.

L'étape suivante – en cours de développement au niveau européen – est de pouvoir échanger plus efficacement des éléments de preuve entre pays. Ainsi, il est aujourd'hui possible de mettre en commun des éléments de preuve dans le cadre d'une équipe commune d'enquête au sein de l'Union européenne (avec le soutien d'Eurojust et d'Europol). De même, l'ordre d'enquête européen permet à un magistrat de demander la réalisation d'opérations d'enquête (comme des perquisitions ou des réquisitions à des opérateurs) dans d'autres pays. Pour être encore plus efficace, dans les cas les plus simples (identification du titulaire d'un abonnement Internet par exemple), des demandes contraignantes pourraient être adressées d'un enquêteur dans un pays A à un opérateur de communications électroniques dans un pays B.

Les enjeux sont aussi normatifs. Ainsi, même si elle n'est pas légalement contraignante, la norme ISO 17025:2005 qui applique l'assurance qualité aux laboratoires d'essais est reconnue au plan européen (en particulier par l'association ENFSI des laboratoires de criminalistique) comme devant s'appliquer dans ces laboratoires. Pour s'appliquer au domaine numérique – manifestement différent des travaux réalisés dans un laboratoire de biochimie par exemple – il a fallu prendre en compte des spécificités quant à la construction et la validation des méthodes. En particulier, il est important de pouvoir régulièrement mettre à jour les logiciels utilisés par les experts judiciaires numériques. Le laboratoire informatique-électronique de l'Institut de Recherche criminelle de la Gendarmerie nationale (IRCGN) en France est accrédité [(2)] pour quatorze natures d'essais différents. D'autres normes trouvent à s'appliquer dans le champ de l'investigation numérique, notamment en matière d'échange d'informations avec les opérateurs. Ainsi, une vingtaine de spécifications techniques [(3)] produites par l'organisation européenne de normalisation ETSI précisent les conditions dans lesquelles le contenu des interceptions judiciaires et les métadonnées sont mis à disposition des autorités par les opérateurs. Ce sont ces standards qui sont mis en œuvre en France dans le cadre de la Plateforme nationale d'Interceptions judiciaires (PNIJ) gérée par le ministère de la Justice.

(2) http://www.cofrac.fr/annexes/sect1/1-1916.pdf
(3) http://www.etsi.org/technologies-clusters/technologies/lawful-interception

Le renseignement criminel

Comme nous l'évoquions en introduction, la collecte et l'analyse de données est réalisée plus globalement dans l'ensemble des enquêtes judiciaires. Ainsi, au-delà des éléments de preuve numérique, ce sont l'ensemble des éléments-clés de l'enquête qui peuvent être transformés en données et exploités, qu'il s'agisse d'informations anthropométriques telles que les collectait Bertillon au XIX^{e} siècle, de détails de l'enquête ou de données issues des analyses d'éléments collectés sur une scène de crime. Bien évidemment, s'agissant de données personnelles liées à une enquête judiciaire, ces traitements font l'objet de textes législatifs et réglementaires les encadrant et d'un contrôle de la CNIL, voire d'un magistrat référent dans certains cas.

Le renseignement criminel traditionnel

Plusieurs types de traitements de données sont mis en œuvre pour réaliser des rapprochements entre enquêtes judiciaires. En France, on trouvera par exemple le traitement des antécédents judiciaires (TAJ) avec les identités des personnes mises en cause et une synthèse des faits incriminés, et pour les phénomènes les plus importants des bases dites sérielles, contenant sur les affaires plus de détails permettant d'opérer des rapprochements plus fins. Au sein d'une même affaire particulièrement complexe, afin de confronter entre eux les indices collectés, d'identifier les contradictions ou simplement de réaliser des synthèses graphiques des relations entre les faits et les acteurs, des outils dits d'analyse criminelle sont mis en œuvre.

Les services de police canadiens font souvent appel à des équipes de recherche universitaire en criminologie pour apporter un regard extérieur sur leurs données : analyse de réseaux, classification des comportements ou encore validation des méthodes d'enquête [4]. En France, ces démarches académiques sont moins courantes, mais on pourra par exemple saluer les travaux du projet MAPAP [5] sur l'analyse de la diffusion des images pédopornographiques *via* les réseaux pair-à-pair (Fournier et Latapy, 2015).

Le renseignement forensique

Plus spécifiquement, les éléments issus des relevés réalisés par les techniciens sur la scène de crime, puis des examens de laboratoires criminalistiques, peuvent être intégrés dans des bases de données pour réaliser des rapprochements. C'est traditionnellement le cas des bases de données biométriques (empreintes digitales, empreintes génétiques), mais c'est aussi possible pour de nombreux autres types d'information moins connus dans le grand public : balistique (traces laissées par une arme sur un projectile), outils (empreinte laissée par les outils utilisés pour ouvrir une porte), lobes d'oreille (qui peuvent laisser une empreinte lorsqu'ils sont collés contre une surface), chaussures ou encore insectes, ADN de cannabis…

Ces bases de connaissances sont parfois nécessaires pour affiner les circonstances d'un fait criminel (par exemple retrouver le lieu géographique auquel correspond une certaine composition de terre) ou encore réaliser des rapprochements entre les faits (identifier le fournisseur d'un produit stupéfiant). La mise en œuvre de telles bases de données forensiques suppose à la fois de combiner des capacités analytiques suffisamment fines et de mettre en œuvre des méthodes de rapprochement efficaces : de nombreux développements sont encore possibles.

(4) On pourra notamment se rapporter aux travaux de l'équipe dirigée par Martin Bouchard de l'Université Simon Fraser de Vancouver.
(5) http://antipaedo.lip6.fr/

La dématérialisation de la procédure pénale

L'évolution suivante du renseignement judiciaire passera par la dématérialisation complète de la procédure pénale depuis la plainte de la victime jusqu'à l'audience devant le tribunal en passant par les pièces de procédure réalisées par les magistrats et les enquêteurs. Aujourd'hui, ces documents circulent essentiellement sous forme imprimée (et souvent signés ou paraphés à chaque page). Le 10 janvier 2018, les ministres de l'Intérieur et de la Justice ont conjointement annoncé [6] le lancement d'un grand projet de dématérialisation de la procédure pénale qui devra notamment permettre à tous les acteurs de la chaîne judiciaire d'accéder en ligne à un dossier unique. Cette mise à disposition de l'ensemble des pièces de l'enquête sous forme numérique ouvre de nouvelles perspectives, tant pour les acteurs de l'enquête judiciaire que pour les avocats des différentes parties.

Vers la maîtrise des données

L'évolution des technologies – notamment celles liées aux méga-données et à l'intelligence artificielle – nous permet d'envisager d'aller beaucoup plus loin dans l'exploitation des données pour la lutte contre la délinquance. Plusieurs projets d'analyse décisionnelle sont ainsi lancés en France (CREOGN, 2017), qui proposent de rapprocher les données issues des enquêtes judiciaires et les données provenant d'autres sources d'information complémentaires (géographiques, sociologiques, économiques, météorologiques, etc.). Elles pourraient ensuite être confrontées aux informations liées à l'action des services répressifs (par exemple la nature, le lieu et la date des opérations de contrôle réalisées) pour permettre d'identifier les mesures les plus efficaces, d'orienter la prise de décision et de favoriser une analyse objective du retour d'expérience.

De nombreux autres axes de travail sont envisagés, pour améliorer le traitement des images et le rapprochement des traces et indices, grâce à l'intégration ou l'amélioration de l'intelligence artificielle. On peut aussi imaginer de faciliter le travail des enquêteurs et des magistrats pour explorer de façon plus systématique l'ensemble des hypothèses probables ou moins probables qu'ils n'auraient pas le temps d'envisager dans un temps raisonnable et leur permettre de n'oublier aucun élément ou piste utile.

La réussite de tels projets ne repose pas que sur des facteurs techniques. Ce travail ne pourra être réalisé que grâce à des échanges de données efficaces avec tous les acteurs concernés (industriels, collectivités locales, chercheurs) : il faut inventer les cadres techniques, réglementaires, éthiques et éventuellement financiers nécessaires à leur mise en place.

Bibliographie

CREOGN – Centre de recherche de l'école des officiers de la Gendarmerie nationale (2017), « Hyperconnexion et résilience », *Revue de la Gendarmerie nationale*, n° 260, pp.146-175.

FOURNIER R. & LATAPY M. (2015), "Temporal Patterns of Pedophile Activity in a P2P Network: First Insights about User Profiles from Big Data", *International Journal of Internet Science*, 10 (1), pp. 8-19.

FREYSSINET É. (2003), « La preuve numérique : Un défi pour l'enquête criminelle du XXIe siècle », *Les Cahiers du numérique*, 4 (3), pp. 205-217.

THIÉRACHE C. & FREYSSINET É. (2018), « La procédure pénale face aux évolutions de la cybercriminalité et du traitement de la preuve numérique », CECyF & Cyberlex.

(6) http://www.justice.gouv.fr/la-garde-des-sceaux-10016/dematerialisation-des-procedures-penales-31168.html

Souveraineté numérique : le rôle des armées

Par le Vice-Amiral Arnaud COUSTILLIÈRE
Directeur général des systèmes d'information et de communication
du ministère des Armées

« *Les armées doivent planifier et conduire les opérations dans l'espace numérique jusqu'au niveau tactique, de façon totalement intégrée, à la chaîne de planification et de conduite des opérations cinétiques.*»

Revue stratégique, §299, octobre 2017

Une souveraineté d'action

Le cyberespace est un milieu artificiel. Il se matérialise par trois couches : une couche physique (serveur, réseaux, terminaux), une couche logique (logiciel, automates, systèmes d'exploitation ou OS) et une couche cognitive (les informations, les liens sociaux). Un objet dans le cyberespace se retrouve dans ces trois dimensions. La dépendance des objets vis-à-vis de ces trois niveaux ne s'oppose pas à la fluidité des échanges. Bien au contraire, les facteurs qui compartimentent les autres espaces comme le temps et la distance sont ici gommés, voire inopérants, l'individu ou l'entreprise pouvant être touchés même au cœur du territoire national. Le milieu devient presque homogène. De cette homogénéité découlent une propriété d'ubiquité, un relatif anonymat et une rémanence. Ainsi, une donnée déposée sur le réseau pouvant se trouver simultanément en plusieurs endroits, il est très coûteux d'en retracer le parcours et illusoire de vouloir en effacer toutes les traces.

Appliquer une notion de souveraineté à un tel milieu n'est pas évident. Une approche consisterait à le comparer aux autres milieux mieux connus et maîtrisés. La souveraineté entendue comme terrestre ou terrienne est une souveraineté d'appartenance. La maîtrise du milieu s'y fait par une occupation permanente du sol. Lorsque l'homme a « conquis » les mers, il a défini une nouvelle notion de souveraineté : une partie de cet espace relève d'une logique de continuité territoriale, mais la haute mer n'appartient à personne, les bâtiments relèvent de leur pavillon. Pour y tenir sa place, l'État y envoie des navires et en assure la surveillance par intermittence : il s'agit d'une souveraineté de présence. Le milieu aérien est quant à lui une prolongation du milieu survolé ; il répond à une souveraineté d'appartenance ou de présence selon les endroits. Mais le cas des aéroports est intéressant lorsqu'on réfléchit à la notion de souveraineté : il illustre un « passage » dont le contrôle doit rester fluide. La souveraineté y est un mélange d'appartenance et de présence complété par une logique d'accès. Enfin, l'espace a sa logique propre puisque la souveraineté spatiale est purement une souveraineté d'accès. L'État est souverain spatialement quand il peut accéder librement à l'espace.

Le cyberespace, quant à lui, n'est à personne mais tout le monde y a accès. Il est matériellement implanté sur un territoire. Aussi, certains États font le choix d'une territorialisation d'Internet afin de mieux contrôler leur cyberespace. C'est le choix de la Russie, de l'Iran ou de la Chine, mais pas celui de la France et des nations occidentales. En comprenant le cyberespace comme un espace de liberté mais aussi de droit, l'État français définit l'ambition de sa souveraineté pour les armées comme la capacité de conduire en propre tout processus numérique d'intérêt national dans le cyberespace, sans y appliquer un contrôle permanent. Cette forme de souveraineté peut être qualifiée de souveraineté d'action.

L'ambition numérique du ministère

L'exercice de la souveraineté se traduit par une stratégie qui s'intègre efficacement dans le déploiement de la politique numérique de l'État. Le rôle des armées dans la cyberdéfense a été présenté dès le Livre blanc de 2008 [1], puis confirmé avec de nouvelles ambitions dans celui de 2013 [2]. Depuis lors, le Premier ministre Manuel Valls a défini les objectifs de la stratégie nationale pour la sécurité du numérique en 2015 [3], le ministre de la Défense Jean-Yves Le Drian a créé un commandement Cyber (COMCYBER) en décembre 2016 [4], soulignant que « l'arme cyber est une arme à part entière, qui fait partie des moyens à disposition du commandement militaire ». Le fait « cyber » a aujourd'hui toute sa place dans la politique de défense de la France. À ce titre, la dimension numérique est présente tout au long de la *Revue stratégique de défense et de sécurité nationale* présentée en octobre 2017 [5] au président de la République.

Par ailleurs, sur un plan plus large, intégrant les opérations mais également le quotidien des personnels et l'amélioration de la relation au citoyen, la ministre des Armées, Florence Parly, a validé en novembre 2017 le document *Ambition numérique* [6] : « La révolution numérique sera un vecteur fort de cette transformation. Je veux la mettre au service du ministère : l'Internet des objets, l'intelligence artificielle ou le *Big Data* sont autant de chantiers ouverts sur lesquels nous devons appuyer le succès de nos armes, l'efficacité et l'excellence dans la conduite de toutes les missions du ministère. »

La révolution numérique en cours est un puissant moteur de transformation et d'accélération de la performance pour toute grande organisation. Le ministère et les armées ont résolument pris le tournant et saisi cette opportunité pour rester à la pointe des meilleures technologies et pratiques.

Cette transformation vise, au travers de nouveaux usages, à s'approprier rapidement et dans les meilleures conditions les technologies émergentes, pour générer des ruptures dans les pratiques, les organisations et les modes de travail ou d'action.

La donnée numérique est au centre de ces enjeux. Il est nécessaire d'apprendre à mieux la traiter, mieux la sécuriser au niveau national et la partager au bénéfice de l'action globale des armées en opérations et dans le fonctionnement quotidien du ministère des Armées.

En tant que pilier de la confiance, l'identité numérique est un autre enjeu d'importance pour l'État. Son déploiement, sa sécurisation et sa viabilisation sont en effet indispensables à la généralisation des démarches dématérialisées, tant dans les relations commerciales que dans la relation du citoyen à une administration moderne.

Toujours dans le cadre de l'ambition numérique du ministère, la souveraineté numérique s'exprime par la volonté du ministère d'amplifier les capacités de développements rapides. Ces capacités favoriseront l'emploi de logiciels aux sources libres, constituant un puissant facteur d'indépendance et d'agilité dans la mise à disposition de nouveaux services.

Au-delà de la technologie, il s'agit aussi de s'attacher de nouveaux talents d'innovateurs et d'injecter agilité et attractivité dans les métiers numériques du ministère. Dans le courant de l'année 2018, la

(1) http://archives.livreblancdefenseetsecurite.gouv.fr/2008/IMG/pdf/livre_blanc_tome1_partie1.pdf
(2) http://www.livreblancdefenseetsecurite.gouv.fr/pdf/le_livre_blanc_de_la_defense_2013.pdf (chap 7.C)
(3) https://www.ssi.gouv.fr/actualite/la-strategie-nationale-pour-la-securite-du-numerique-une-reponse-aux-nouveaux-enjeux-des-usages-numeriques/
(4) http://discours.vie-publique.fr/notices/163003632.html
(5) https://www.defense.gouv.fr/dgris/presentation/evenements/revue-strategique-de-defense-et-de-securite-nationale-2017
(6) https://www.defense.gouv.fr/actualites/articles/innovation-numerique-le-ministere-poursuit-sa-transformation

création de la DGNUM [7] comme chef d'orchestre de cette dynamique permettra d'accélérer et de structurer cette démarche de profonde modernisation et de transformation menée en cohérence avec celle initiée en interministériel au sein d'Action Publique 2022 [8]. Trois objectifs stratégiques de performance structurent cette démarche :

- Garantir la supériorité opérationnelle et la maîtrise de l'information sur les théâtres d'opérations ;
- Renforcer l'efficience des soutiens et faciliter le quotidien des personnels ;
- Améliorer la relation au citoyen et l'attractivité du ministère.

L'enjeu de la connaissance : anticipation et renseignement

Action aux canaux multiples, l'anticipation repose notamment sur le renseignement, tant dans sa finalité d'action que dans son acception de veille stratégique. Sur le long terme, il s'agit en effet d'anticiper les évolutions et les tendances pour permettre à la France de décider et d'agir de manière autonome et souveraine. C'est le rôle de la veille stratégique conduite au sein du ministère des Armées.

Sur le temps plus court et potentiellement à fin d'action, le besoin en renseignement est déterminant. Théoriquement, une ligne de séparation existe entre le renseignement d'origine cyber (ROC) et le renseignement d'intérêt cyber (RIC). Si le ROC est défini comme le renseignement provenant de sources cybernétiques (sources ouvertes, essentiellement mais pas toujours de l'Internet, recherches informatiques et investigations de supports numériques comme les disques durs, les clés USB, les téléphones, les tablettes ou l'électronique des systèmes d'armes…), le RIC a quant à lui pour vocation d'apporter à la chaîne de commandement opérationnel de la cyberdéfense militaire les informations dont la connaissance et la compréhension sont nécessaires pour opérer en sécurité dans le cyberespace. Il vise à évaluer la menace cyber qui pèse sur les forces en opérations et à exploiter les opportunités dans le camp adverse.

La capacité à échanger le renseignement est essentielle, tant à l'intérieur du ministère qu'avec nos alliés, tout comme celle d'orienter les recherches. Ce processus a lieu autour du cycle classique du renseignement : orientation, recherche et acquisition, exploitation, diffusion.

La capacité d'action

En dernier ressort, la souveraineté doit pouvoir être garantie par une capacité d'action. Deux impératifs sont corollaires de la capacité d'action : la résilience des systèmes des armées et la possession d'options défensives comme offensives.

Un certain zèle peut conduire à confondre la résilience nécessaire à la souveraineté avec une indépendance ou une autonomie totale des moyens. Dans un système économique mondialisé, cette orientation n'apparaît pas pertinente, les armées n'étant pas en mesure de contrôler de bout en bout l'autonomie de leur production en électronique et en informatique. Partageant sur ce point la position d'autres partenaires, le choix est fait de ne fortement sécuriser que ce qui est vital au sein d'un environnement qu'on estime peu sûr. Les armées n'ont ainsi besoin de maîtriser en propre que quelques composants bien précis pour pouvoir sécuriser un ensemble composé de briques peu fiables. Tout d'abord, il s'agit d'avoir des outils de cryptographie souverains pour assurer l'intégrité et la confidentialité des données. Ensuite, la maîtrise des réseaux nécessite notamment la possession de sondes de détection entièrement fiables et maîtrisées, afin de garantir la disponibilité des données. Enfin, il faut des algorithmes nationaux pour assurer le traitement de ces données.

(7) Direction générale du Numérique et des Systèmes d'Information et de Communication, ex-DGSIC.
(8) https://www.economie.gouv.fr/lancement-programme-action-publique-2022

La garantie de disposer des moyens présentés répond aux besoins fondamentaux des armées. En effet, l'absence de certification systématique des matériels et des logiciels est palliée par le chiffrement et la disponibilité de service. Par ailleurs, bien que les armées ne soient pas autonomes en matériels et en logiciels, elles peuvent être considérées comme indépendantes par la diversité des sources d'approvisionnements et de fabrications, et l'emploi de plusieurs réseaux protégés et physiquement séparés.

En outre, croire que les composants ou les logiciels pourraient tous être nationaux ou européens est une illusion. Cet effort serait inutile car, même développés en propre, les produits numériques ne pourront jamais être considérés comme parfaitement fiables. De plus, leur maintien à jour et en condition de sécurité est bien souvent un défi coûteux, rendant les logiciels propriétaires non nationaux très attractifs.

Enfin, le développement des *data sciences* et du *machine learning* crée un nouveau besoin : posséder des ensembles, ou *sets* nationaux de données labellisées pour entraîner ou évaluer les algorithmes [9]. Par exemple, dans un environnement de plus en plus dépendant du renseignement en source ouverte et des moteurs de recherche, il est important de pouvoir évaluer la fiabilité et les biais des moteurs publics. Ce besoin dépasse le cadre strict des armées et interroge sur la confiance à accorder à des algorithmes non publics. La validation des algorithmes privés par des *sets* secrets de données gérés par des organismes de contrôle publics pourrait répondre à ce besoin de certification.

> « L'enjeu considérable que représente la menace cyber appelle un renforcement substantiel à la fois des moyens défensifs et offensifs de la France. La capacité de détection et d'attribution des attaques, qui repose sur l'acquisition de renseignement d'origine humaine aussi bien que technique, en sera un élément-clé. »
>
> (*Revue stratégique*, §299, octobre 2017)

Le second aspect de la capacité d'action est de conserver une force de frappe défensive ou offensive prête à servir. Un cyberespace omniprésent est en effet un enjeu stratégique tant civil que militaire. La communauté internationale éprouve d'ailleurs des difficultés à faire émerger un cadre juridique international adapté. Outre les grandes puissances et certaines puissances moyennes, qui disposent de l'ensemble des capacités d'actions cyber et qui, dans le cadre de concurrences politique ou économique, peuvent conduire des actions de renseignement inamicales, de nombreuses nations ou organisations structurées peuvent conduire des opérations dépassant le stade de l'espionnage. Les groupes activistes doivent également être pris en compte, ainsi que ceux liés au terrorisme international et aux différentes formes de nationalismes.

Le cyberespace est devenu un champ de confrontation à part entière. Les armées doivent disposer des capacités de maîtrise et d'action dans ce milieu. De même, l'État, au travers de ses institutions, doit maintenir son fonctionnement, les activités d'importance vitale, la sécurité et l'activité économique face à ces menaces cybernétiques. Elles peuvent être regroupées en trois familles : celles ciblant directement les informations, celles visant les systèmes d'information eux-mêmes et celles qui passent par le cyberespace pour viser des objectifs physiques (équipements et installations critiques, etc.).

(9) Les modèles supervisés sont entraînés avec des données dont les résultats attendus, « labels », ont été complétés auparavant (manuellement ou par des automates). Les biais de la labellisation du set d'entraînement se retrouvent ainsi « gravés » dans le fonctionnement de l'algorithme.

Or, dans le cyberespace, les notions de preuve, d'imputabilité des actions ou attributions sont plus complexes à faire émerger. Cela modifie le rapport de force classique entre l'attaquant et le défenseur : l'avantage relatif qu'a l'attaquant est renforcé par la prolifération des technologies d'attaque, comme par exemple Wannacry [10]. L'action dans le cyberespace est délicate à conduire, et la France se veut moteur dans la construction d'un cadre international mettant en place des processus, pour les volets tant défensif qu'offensif.

Le ministère des Armées a un rôle bien précis pour ce qui concerne la souveraineté numérique nationale : garantir à la fois un haut niveau de performance dans toutes ses composantes en bénéficiant de la révolution numérique, mais aussi se déployer et combattre dans ce nouvel espace. Pour garantir de façon pérenne leur réussite, les armées sont en pleine transformation digitale, tant de leurs processus que de leurs équipements. La création du COMCYBER puis de la DGNUM dans le courant de l'année 2018 démontre que le ministère a adapté son organisation pour répondre aux défis posés par l'espace numérique, milieu évolutif, compressé et fait d'immédiateté, qui interroge l'ensemble de nos modèles.

(10) Logiciel malveillant de type ransomware auto-répliquant.

Big Data : données sur les entreprises et marketing prédictif B2B

Par François BANCILHON
Sidetrade

Quoi de neuf sur les données ?

Le Big Data, c'est la conjonction de trois phénomènes : la disponibilité de données nouvelles et en grande quantité, la disponibilité de machines de plus en plus puissantes (la loi de Moore continue à faire des siennes) et la disponibilité de technologies de parallélisation qui permettent de traiter ces données massives grâce à ces machines puissantes.

Les données nouvelles sont essentiellement venues des réseaux sociaux, de l'Internet des objets, des données générées par le web et de l'open data. Les deux premières sources de données étant largement connues, nous détaillons uniquement les deux dernières.

Les données du web qui ont contribué au Big Data sont de deux types : les données du web elles-mêmes et celles liées au comportement des utilisateurs du web. La capacité à recueillir des données sur le web par *crawling* (aspiration de l'ensemble du contenu d'un site web sans essayer de le comprendre) ou par *scraping* (recueil du contenu d'un site web en en connaissant la structure, connaissance utilisée pour extraire des informations structurées) permet à de nombreux opérateurs de construire des jeux de données riches et importants à partir du web. C'est cette activité qui fait que 52 % du trafic Internet soit généré par des robots plutôt que par des humains. La capacité à recueillir les données de comportement des utilisateurs du web a conduit à la mise en place, par les gestionnaires de sites web, de plateformes DMP (*data management platforms*) qui mémorisent les comportements des usagers du web (qui est venu sur mon site, quand et pour y voir quoi) et qui étendent cette information à des ensembles de sites.

L'open data (ouverture des données) est la mise à disposition des entreprises (pour réutilisation) et des citoyens (pour consultation et par besoin de transparence) des données dont se sert la puissance publique dans l'exercice de ses fonctions. Échappent bien entendu à cette ouverture tout ce qui relève du « secret défense » et les données personnelles. Le mouvement open data, lancé par l'administration Obama à la fin des années 2000, a trouvé un écho fort dans l'ensemble des pays démocratiques, à des degrés divers. Nous verrons plus bas sa traduction en France.

Pour traiter ces données massives, des technologies ont été développées, qui permettent à tous, de la plus petite *start-up* au plus grand groupe, de recueillir massivement des données, de les traiter et de construire applications et nouveaux usages pour les exploiter.

Le Big Data, basé sur ces données et la capacité à les exploiter, s'est donc développé dans de nombreux domaines : transport, médecine, ville intelligente, publicité, agriculture, marketing *Business-to-Consumer* (B2C), etc. *A priori*, aucun domaine ne devrait y échapper.

Cet article se concentre sur le sujet du marketing Business-to-Business (B2B).

Quoi de neuf pour les données sur les entreprises ?

Un grand nombre de données officielles sur les entreprises, qui n'étaient pas visibles, sont tout à coup devenues disponibles à tous grâce à l'open data. C'est le cas des données émises par les entreprises sur elles-mêmes et de celles produites à propos d'elles par d'autres (presse et médias Internet).

Les entreprises communiquent (et donc se décrivent) sur leurs sites web, leurs réseaux sociaux, leurs applications mobiles, leurs sites éventuels de commerce électronique et les sites de recrutement sur lesquels elles publient des offres de postes. Nous verrons plus loin le taux d'équipement des entreprises en chacun de ces outils. Par ailleurs, les médias Internet communiquent régulièrement sur les entreprises, ce qui est une deuxième source d'information.

Quoi de neuf en France sur les données sur les entreprises ?

Quelles données open data sont disponibles en France sur les entreprises ? Quelle communication ces mêmes entreprises font-elles sur le web ?

L'ouverture des données

Le mouvement d'ouverture des données, amorcé en France à partir de 2010, s'est fortement accéléré sous l'impulsion d'Etalab à partir de 2013. De nombreux lois et décrets, sous les divers gouvernements, ont réaffirmé le principe de l'open data par défaut (qui n'est pas complètement appliqué, mais progresse régulièrement). Après l'État, les collectivités se sont emparées de l'idée. Ce mouvement s'est notamment traduit par de nombreuses ouvertures de données sur les entreprises. Le fichier SIRENE (Système national d'identification et du répertoire des entreprises et de leurs établissements), géré par l'Insee, a été ouvert au début de 2017. Le RNCS (Registre national des Commerces et des Sociétés), après un long combat d'arrière-garde d'Infogreffe pour en empêcher l'ouverture, a été ouvert début 2018. C'est finalement à l'INPI qu'il est revenu de mettre à disposition ces données, collectées par les greffes des tribunaux de commerce. L'INPI avait déjà en 2016 ouvert ses données (marques, brevets et dessins). Le Bodacc (Bulletin officiel des Annonces civiles et commerciales) avait lui aussi été rendu disponible à la même date par la Direction de l'Information légale et administrative (DILA). Le ministère de la Recherche et de l'Enseignement supérieur a contribué à l'édifice en publiant la liste des entités publiques de recherche et des projets coopératifs entre recherche et industrie, ce qui s'est traduit par la mise en place de l'application scanR qui offre l'accès à ces données. Le RNA (Registre national des Associations) est lui aussi ouvert depuis 2017.

Ainsi, en un intervalle de deux ans, c'est l'ensemble des données détenues par l'État sur les entreprises qui est devenu disponible en open data. Avant cette ouverture, la plupart de ces données n'étaient disponibles que de façon payante : pour l'ensemble de ces données, il fallait compter plusieurs centaines de milliers d'euros d'abonnement mensuel pour y accéder.

Le premier impact de cette ouverture a été l'accès immédiat des *start-ups* qui ont pu offrir des services innovants à partir de ces données. Le deuxième impact est l'amélioration de la qualité de ces données par le retour de ceux qui y ont accès.

Maturité numérique des entreprises

Les entreprises françaises participent au mouvement général d'accès au web. Elles disposent de sites web, elles sont présentes sur les réseaux sociaux, elles sont présentes sur Internet *via* des applications mobiles. Le tableau ci-après donne le taux d'équipement en sites web, sites e-commerce, Facebook, Twitter et LinkedIn. Pour les entreprises françaises, nous n'avons retenu que les « sociétés commerciales » (catégorie Insee).

Pays	Entreprises	Sites Web	ecommerce (% de SW)	twitter	FB	linkedIn
UK	4,260,628	23.89%	7.59%	8.29%	7.76%	3.66%
FR	2,454,396	25.20%	9.64%	4.06%	9.22%	1.80%
BE	1,700,340	6.69%	5.87%	0.73%	2.61%	0.50%
NL	2,760,265	30.64%	6.63%	5.34%	9.36%	2.68%

Le taux d'équipement croît assez vite dès qu'on considère des entreprises d'une certaine taille : si l'on se restreint aux entreprises d'un chiffre d'affaires d'au moins 1M€, le taux d'équipement en sites web est de 60 % pour la France.

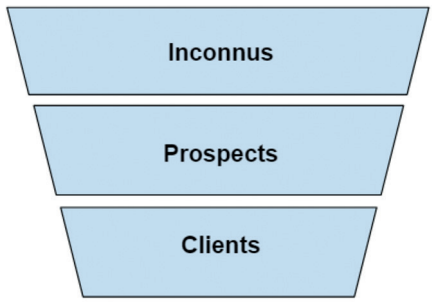
Quel impact sur le marketing B2B ?

On peut diviser le marketing B2B en deux grandes catégories d'activités : la gestion de l'« entonnoir commercial », processus central du marketing opérationnel allant de l'identification du prospect à sa transformation en client, et les études de marché qui sont des études générales pour comprendre un ou plusieurs marchés et leur évolution.

La gestion de l'entonnoir

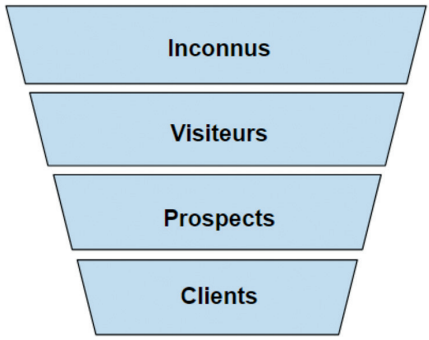
La première des figures ci-dessous illustre l'entonnoir classique de la démarche marketing. Il s'agit d'identifier parmi l'ensemble des entreprises celles qui peuvent devenir clientes, et de les transformer en prospects avec lesquels on établit une relation commerciale, puis qu'on transforme en clients. Le passage du statut d'inconnu à celui de prospect se faisait jusqu'à maintenant par des méthodes traditionnelles : publicité, présence sur des salons, publipostage, télémarketing et plus récemment mailing de masse à partir de listes d'adresses acquises auprès de vendeurs de fichiers.

La seconde figure illustre le nouvel entonnoir marketing dans lequel une partie importante de l'activité se passe sur le site web, ce qui justifie l'apparition d'une nouvelle catégorie, celle des visiteurs de ce site.



Qui sont les entreprises ou les personnes à chaque étage de l'entonnoir ?

- Les inconnus : ce sont les 2,5 M d'entreprises que vous ne connaissez pas, qui sont pour certaines d'entre elles des clients potentiels de vos produits ou services ;
- Les visiteurs de votre site web sont de deux catégories :
  - non identifiés : ils sont venus sur votre site, attirés par votre *inbound marketing*, vous ne les connaissez pas, mais vous les voyez circuler sur vos pages et vous mémorisez leur parcours ;
  - identifiés : ils sont venus sur votre site, attirés par votre *inbound marketing*, ils se sont inscrits (pour avoir un livre blanc, une démonstration ou un autre contenu), donc vous les connaissez, vous les voyez circuler sur vos pages et vous mémorisez aussi leur parcours ;



- Les prospects : ils ont parlé avec vos commerciaux ou interagi avec votre système ;
- Les clients : ils utilisent vos produits ou services, vous voyez ce qu'ils en font ou ce qu'ils n'en font pas.

Le jeu de l'entonnoir consiste à faire descendre le plus possible d'entreprises ou de personnes dans l'entonnoir et à avoir un taux de transformation entre l'étage n et l'étage n-1 le plus élevé possible. Plus précisément :

- Les inconnus : il s'agit de trouver parmi eux ceux qui sont vos clients potentiels (donc ceux qui ont besoin de vos produits ou services) et de les faire venir sur votre site par du marketing *in-bound* (à partir de contenu ou de référencement) ou *out-bound* (par des mailings de masse) ;
- Les visiteurs non identifiés : vous pouvez en identifier certains (par *IP-tracking* qui consiste à rattacher l'adresse IP de votre visiteur à son entreprise), vous pouvez apprendre des informations supplémentaires sur eux (*via* une *data management platform* qui mémorise leurs trajets sur d'autres sites), vous pouvez aussi les convaincre de s'identifier pour bénéficier d'informations supplémentaires ;
- Les visiteurs identifiés : il s'agit de les transformer en prospects ;
- Les prospects : il s'agit de les transformer en clients ;
- Les clients : vous suivez leur comportement, vous voulez les convaincre d'acheter plus (*up-sell*), d'acheter d'autres produits de votre gamme (*cross-sell*) et de ne pas vous quitter (rétention).

Voyons ce que nous avons comme données sur les éléments à chaque étage de l'entonnoir :

- Les inconnus : l'ensemble des données sur les entreprises ;
- Les visiteurs non identifiés : leur comportement sur votre site, et éventuellement sur d'autres sites ;
- Les visiteurs identifiés : leur comportement sur votre site, et éventuellement sur d'autres sites, ainsi que les données sur leur entreprise ;
- Les prospects : leur comportement sur votre site, sur d'autres sites ainsi que les données sur leur entreprise et l'historique de votre relation commerciale avec eux ;
- Les clients : leur comportement sur votre site, et éventuellement sur d'autres sites ainsi que les données sur leurs entreprises et l'historique de votre relation avec eux et de leur usage de votre produit.

Voyons maintenant avec quelles technologies on peut utiliser ces données pour accomplir notre objectif de faire descendre vos interlocuteurs dans l'entonnoir :

- Les inconnus : outils de filtrage, de recherche et de *scoring* par *machine learning ;*
- Les visiteurs non identifiés : *IP tracking*, *scoring* par *machine learning*, *marketing automation ;*
- Les visiteurs identifiés : *scoring* par *machine learning*, *marketing automation ;*
- Les prospects : *scoring* par *machine learning ;*
- Les clients : *scoring* par *machine learning.*

En quoi consiste ce *machine learning* pratiqué à chaque étage de l'entonnoir ? À partir des données accumulées sur le comportement de l'entonnoir dans le passé, on construit un modèle prédictif plus ou moins sophistiqué qui nous permet d'identifier les candidats les plus aptes à être transformés.

Le tableau ci-dessous résume ce que nous venons de décrire.

Cibles	Objectif	Données	Technologies et outils
Inconnus	Identifier et engager	Données externes sur les entreprises	*Machine learning*
Visiteurs inconnus	Identifier et engager	Comportement sur le site	*IP Tracking* et *Machine learning* *Marketing automation*
Visiteurs connus	Transformer en prospects	Données externes sur les entreprises + comportement sur le site	*Machine learning* *Marketing automation*
Prospects	Transformer en clients	Données externes sur les entreprises + comportement sur le site + interaction commerciale	*Machine learning* CRM
Clients	Rétention, *up-sell, cross-sell*	Données externes sur les entreprises + comportement sur le site + interaction commerciale + utilisation du produit/service	*Machine learning* CRM

Les études d'ensemble

Sous cette catégorie, nous rangeons l'évaluation de la taille d'un marché et de son évolution, l'identification d'un écosystème et la segmentation d'un marché par type d'activité. Ce type d'étude permet par exemple d'identifier des entreprises innovantes ou de comprendre le basculement de la technologie du moteur à explosion du diesel vers l'essence. Il s'agit donc de qualifier ou de quantifier un phénomène. Pour faire ce type d'étude, on peut en passer commande à un cabinet de conseil. Le cabinet va identifier entre vingt et cent entreprises qu'il considère comme représentatives du secteur. Il va préparer un questionnaire, l'administrer à son panel, produire un rapport qualitatif et quantitatif, et faire des recommandations. Le rapport sera imprimé et distribué aux clients.

Ce qui caractérise l'approche traditionnelle de ce type d'étude est que le rapport est en papier (une version électronique en pdf sera certainement disponible mais cela ne change pas le problème), qu'il donne une photographie à un instant déterminé mais peut être dépassé six mois plus tard et qu'il se fonde sur un échantillon.

L'ensemble des données disponibles sur les entreprises en temps réel, la capacité à les maintenir à jour automatiquement dans le temps, la possibilité de développer des outils interactifs de visualisation, la capacité à maintenir des données sur l'ensemble des entreprises et non pas sur un échantillon, permettent désormais d'adopter une approche complètement différente, avec un outil :

- en grandeur nature : on ne prend pas un échantillon, mais l'ensemble réel des entreprises concernées ;
- en temps réel : on ne fait pas une photo, mais on suit l'évolution en temps réel des données ;
- interactif : ce n'est pas un rapport mais une application interactive qui permet de voir divers points de vue.

Ces tableaux de bord dynamiques, en temps réel, interactifs et exhaustifs, ne sont pas réellement concurrents des autres méthodes, plus qualitatives, ils tendent plutôt à les compléter.

Le Big Data est en train de changer la plupart, voire la totalité des activités économiques et sociales. S'il a pénétré un peu plus tardivement le domaine du marketing B2B, il est en train de bouleverser la façon d'aborder ce marketing. De nouvelles données et de nouvelles technologies fournissent des outils nouveaux qui offrent aux entreprises des résultats très fortement améliorés. Ils les aident à acquérir de nouveaux clients et à augmenter leurs revenus avec leurs clients existants. Ils leur fournissent une vision détaillée et en temps réel de leurs marchés.

Les apports des nouvelles technologies numériques pour la maintenance et l'exploitation du parc nucléaire d'EDF

Par Grégoire MOREAU
Direction Recherche et Développement d'EDF
Bruno SUTY
Directeur des Systèmes d'Information et de la Transition numérique industrielle de la Direction du Parc Nucléaire et Thermique d'EDF
et Vincent PERTUY
Architecte d'entreprise de la Direction du Parc Nucléaire et Thermique d'EDF

Le contexte du parc nucléaire d'EDF

Précisons tout d'abord quelques éléments de contexte relatifs au parc électronucléaire exploité par EDF. Ce parc est constitué de cinquante-huit réacteurs, de même technologie connue sous l'appellation « Réacteur à Eau sous Pression » (REP) issue de la filière « Pressurized Water Reactor » (PWR) conçue aux États-Unis. L'âge moyen des réacteurs est de trente-deux ans. La construction de la majorité d'entre eux a été décidée après la survenance de la crise pétrolière de 1973 : il s'agissait alors du programme électronucléaire français. Pour des questions de productivité, ces réacteurs ont été conçus par séries appelées « paliers » (cf. figure 1). Les réacteurs d'un même palier sont standardisés dans leur *design* comme pour le choix des matériels. On est donc en présence de flottes de matériels identiques fonctionnant sur différents réacteurs simultanément. Un modèle cohérent de données de description des installations ayant été adopté dès l'origine, les inter-comparaisons sont possibles et particulièrement pertinentes pour optimiser la performance de l'exploitation.

Figure 1 – Parc nucléaire en exploitation d'EDF. ©EDF

Le second élément de contexte concerne les enjeux de ce parc. En tout premier lieu, il y a un impératif de sûreté associé à l'exploitation électronucléaire et qu'on retrouve à plusieurs niveaux : choix de conception, règles d'exploitation et également implication managériale. Une des clés de voûte du management de la sûreté est le retour d'expérience. Chaque événement qui survient doit être méthodiquement analysé pour éviter qu'il ne se reproduise. Cette exigence impose de conserver toutes les données et ceci est valable pour toutes les phases de vie de ces cinquante-huit réacteurs : conception, construction, exploitation puis déconstruction. Par ailleurs, la performance de ce parc est pour partie liée à la capacité

de l'exploitant à optimiser les volumes de maintenance car ils impactent directement la durée des arrêts de réacteur.

De ces éléments de contexte résulte l'existence pour le parc nucléaire d'un gros volume de données disponibles et ordonnées, propices à la mise en œuvre des techniques naissantes du « Data Analytics ».

Concrètement, ces nouvelles techniques permettent de mener à bien des analyses dont des exemples, ou cas d'usages, sont décrits plus bas. Ces analyses n'étaient auparavant réalisées qu'au prix d'efforts conséquents de collecte et consolidation des données d'historique relatives au problème à traiter à partir de sources distinctes.

Il est utile à ce stade de préciser que ces analyses sont grandement facilitées par la mise en place d'une infrastructure de type Big Data, permettant de stocker et traiter des données massives et hétérogènes. La Direction du parc nucléaire d'EDF a décidé d'investir dans ce type d'infrastructure en 2015. À l'horizon du projet, EDF disposera pour optimiser son exploitation d'un *data lake* de type Hadoop comportant une grande variété de données numériques, textuelles, structurées ou non. Il s'agira principalement et de manière non exhaustive :

- de données de *process* : séries temporelles issues des capteurs présents sur l'installation (pression, température, débit, niveau, vibration…), relevés de rondes d'exploitation, relevés issus de la surveillance chimique et radiochimique des circuits ;
- de données de contexte issues de la GMAO (Gestion de Maintenance assistée par Ordinateur), de bases d'analyses d'événements d'exploitation, des progiciels de planification et de consignation [(1)] des installations.

Les cas d'usages opérationnels déployés ou en cours de déploiement et les techniques d'analyse mises en œuvre

Le premier usage, qui pourrait sembler d'abord trivial, concerne la visualisation des données. En effet, agréger de façon claire, intelligible et rapide de gros volumes de données grâce à des outils de « data visualisation » performants constitue le premier besoin commun à l'ensemble des utilisateurs. Les nouveaux outils permettent en particulier de visualiser aisément des données hétérogènes. Ceci améliore par exemple la réactivité des personnes en charge de surveiller et diagnostiquer l'état des matériels. En effet, lors de la détection d'une évolution des paramètres de fonctionnement du matériel, ce type d'outil leur fournit une vision consolidée des sollicitations subies, des opérations de maintenance réalisées et du retour d'expérience d'évolutions comparables déjà survenues sur les matériels analogues du parc.

Un deuxième usage concerne les transitoires d'exploitation, c'est-à-dire les périodes pendant lesquelles on conduit l'installation d'un état de référence caractérisé par des variables physiques stables vers un autre état de référence. La bonne maîtrise de ces transitoires est un enjeu important pour l'exploitant. En effet, un transitoire est une phase d'exploitation qui augmente la probabilité d'atteindre un seuil de protection de l'installation, le plus redouté étant l'« Arrêt Automatique Réacteur » qui conduit à une mise à l'arrêt très rapide ; il est donc potentiellement pénalisant en termes de sûreté et de production. Par ailleurs, les transitoires sollicitent les matériels et ont donc une incidence sur leur durée de vie, ce qu'on vérifie sur l'exemple des circuits soumis à pression : lors de leur conception, ils sont dimensionnés pour pouvoir subir en service un certain nombre de situations de pression et température correspondant à un chargement mécanique, mais il convient,

(1) La consignation est une suite chronologique d'opérations indispensables et réglementées sur les installations qui permettent d'assurer la sécurité du personnel pendant son intervention.

en exploitation, d'effectuer une comptabilisation de ces situations de pression et de température occasionnées lors des transitoires. Optimiser la conduite de l'installation avec une optique de durée de vie des matériels et de sûreté passe donc par des analyses *a posteriori* des transitoires. Il s'avère que celles-ci sont grandement facilitées par les outils de « Data Analytics » qui permettent d'examiner rapidement de gros volumes de données de *process* en identifiant les transitoires.

Un troisième usage concerne la maintenance prédictive des matériels. Dans un contexte économiquement contraint, ce type de maintenance repose sur une prévision de la dégradation des matériels qui permet de planifier les actes de maintenance de façon optimale avant la survenue de leur défaillance. Lorsque le retour d'expérience sur le matériel est important et qu'un certain nombre de défaillances ont été observées, une étude statistique permet d'établir une loi de fiabilité du matériel. Une telle étude nécessite un travail préalable indispensable de recueil, de prétraitement et de mise au format des données, travail qui a longtemps constitué un frein souvent rédhibitoire pour la mise en œuvre de ces approches. Là encore, quelques cas d'usages traités ont montré que l'emploi du « Data Analytics » facilitait grandement ce type d'étude et qu'il permettait, grâce au croisement aisé des différentes bases de données, d'établir des lois de fiabilité intégrant les facteurs d'influence que sont les conditions d'usage du matériel (durées de fonctionnement, cumul de manœuvres, de mises en service ou de mises à l'arrêt, conditions d'ambiance, caractéristiques chimiques du fluide véhiculé, etc.). Le déploiement généralisé de ces méthodes va permettre d'actualiser plus fréquemment les programmes de maintenance préventive en définissant des périodicités de maintenance des matériels en fonction de leur usage. Par ailleurs, à l'heure où le parc nucléaire est engagé dans un programme sans précédent pour prolonger la durée de fonctionnement de ses réacteurs au-delà de quarante ans, ces approches vont permettre d'optimiser les programmes de remplacement des gros composants en y établissant des priorités, ce qui engendrera des gains financiers.

Ces quelques usages du « Data Analytics » ne représentent qu'une partie du potentiel de valorisation des données pressenti sur le parc nucléaire. Sur les autres filières de production, des perspectives existent également. Pour saisir ces opportunités, EDF dote l'ensemble de ses filières de production d'une entité commune dénommée « Usine Data Analytics pour la production ». Cette entité transverse est le fruit d'une réflexion menée par les directions métiers de production, par la filière Système d'Information et par la direction Recherche et Développement du groupe EDF. Elle regroupe des moyens techniques et humains permettant d'accélérer le passage du concept à l'application sur le terrain. Pour les parcs de production d'EDF, elle capitalise et mutualise les outils, méthodes et usages relatifs aux « Data Analytics ».

Explicitons à présent les principales techniques d'analyse de données mises en œuvre.

Les données issues des capteurs présents sur le *process* et rafraîchies en temps réel conduisent à mettre en œuvre des traitements très rapides de gros volumes de séries temporelles. Ceci est utilisé par exemple pour comptabiliser les situations de fonctionnement pénalisantes pour un matériel donné, ou bien pour établir une classification des modes de fonctionnement d'un matériel à partir de l'analyse du passé.

Le traitement de grandes quantités de comptes-rendus d'intervention issus de la GMAO et des bases de retours d'expériences nécessite d'exploiter les techniques de traitement automatique du langage (TAL), connues également sous le vocable de *text mining*. Ce sont ces techniques qui permettent par exemple d'obtenir une information structurée à partir d'un nombre important de comptes-rendus formalisés par des intervenants différents dans un champ en texte libre (exemple : fréquence de remplacement d'une pièce d'usure) et qui enrichissent les études de fiabilité.

Des techniques plus avancées sont à l'étude pour améliorer les systèmes d'*e-monitoring* (surveillance à distance) déployés sur le parc nucléaire. Ces systèmes sont basés sur une comparaison en temps réel des paramètres d'état du matériel avec les domaines de fonctionnement connus et issus

d'un apprentissage sur ce dernier. Ils permettent aujourd'hui de détecter précocement des anomalies avant défaillance sur un matériel en fonctionnement. Les améliorations visées portent sur l'établissement, presque en temps réel, d'un diagnostic, d'un pronostic et d'une aide à la décision après détection de l'anomalie. Les pistes explorées pour réaliser ces progrès passent par des algorithmes innovants de traitement de données massives et hétérogènes, ou encore par un couplage avec des modélisations physiques des composants permettant d'estimer l'évolution de l'anomalie, et donc la durée de vie résiduelle du matériel lorsque les données de défaillance sont peu nombreuses.

Enfin, les *data lakes* offrent l'opportunité d'utiliser les réseaux de neurones profonds pour modéliser des phénomènes complexes pour lesquels l'emploi des sciences de l'ingénieur n'a pas apporté de solutions satisfaisantes. Plusieurs études de ce type sont actuellement menées.

Des exemples de travaux relatifs au « patrimoine data » du parc nucléaire

Comme expliqué précédemment, l'exploitation des données avec ces techniques émergentes en cours de déploiement va permettre au parc nucléaire d'EDF de réaliser des gains de performance significatifs. Cependant, cela ne doit pas occulter les efforts qui doivent être déployés pour augmenter et améliorer le potentiel intrinsèque du « patrimoine de données ».

Cette préoccupation est majeure dans le métier d'exploitant nucléaire. Pour l'illustrer, présentons ici deux chantiers majeurs d'amélioration menés dans ce domaine.

Le premier de ces chantiers est achevé, il concerne la structure des données. Bien que le parc nucléaire ait été construit comme indiqué plus haut par paliers, EDF ne tirait pas tous les bénéfices de la standardisation du fait de la non-prise en compte de l'effet de parc ou de palier dans les systèmes de GMAO utilisés jusqu'en 2010. Toute évolution du référentiel d'exploitation devait par exemple être déclinée séparément au niveau du système d'information de chaque centrale nucléaire. La rénovation du système d'information du nucléaire (Programme « SdIN ») entreprise depuis 2009 a été mise à profit pour résoudre ce problème. Concrètement, des évolutions fonctionnelles du progiciel d'*asset management* retenu ont été réalisées pour prendre en compte cette

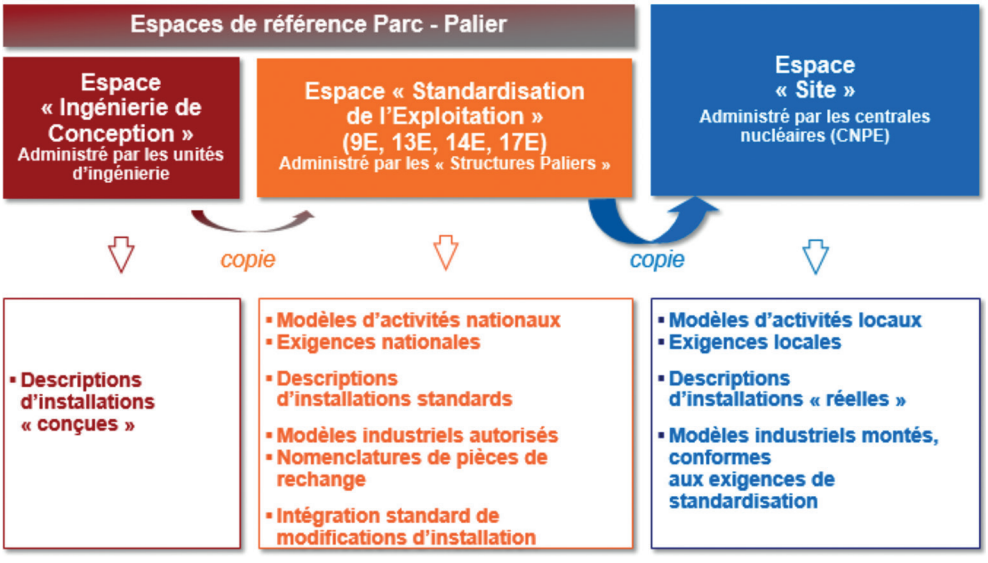


Figure 2 – Espaces des données de la GMAO des installations nucléaires d'EDF. ©EDF

particularité de son parc. Des données de référence techniques « palier » ont ensuite été créées : modèles de matériel installé, modèles d'exigence, modèles d'activité. Enfin, des structures spécifiques de gouvernance et de mise à jour de ces données ont été mises en place pour chaque palier (« Structures Paliers »). La figure 2 représente la logique des espaces d'informations retenue.

Le deuxième de ces chantiers est en cours et est relatif à la numérisation et à la collecte des données au plus proche du terrain. L'apparition des tablettes rend en effet possible la mise en œuvre de dossiers d'intervention électronique dénommés « e-DRT » (Dossier de Réalisation de Travaux électroniques) pour les opérations de maintenance et d'exploitation sur le terrain (cf. figure 3).

Cette évolution, en cours d'expérimentation sur 5 des 19 sites du parc nucléaire, nécessite un travail important de mise sous format électronique des procédures de travail. Elle apporte de multiples avantages :

- simplification du travail de l'intervenant avec un dossier plus lisible (gains sur la non-qualité de maintenance et donc sur la sûreté) ;
- amélioration du temps « métal » (temps uniquement consacré aux gestes techniques de l'intervention de maintenance), aide en ligne à l'intervenant depuis la tablette en cas d'événement non prévu, suivi de tendance en temps réel sur les relevés de paramètres, prise de photos…
- pas de double saisie des données relevées sur le terrain, archivage simplifié ;
- communication plus réactive sur l'avancement des activités vers l'équipe de pilotage et de coordination de l'ensemble des opérations de maintenance réalisées à l'occasion d'une mise à l'arrêt du réacteur (gain de productivité du fait de l'optimisation des enclenchements entre opérations).

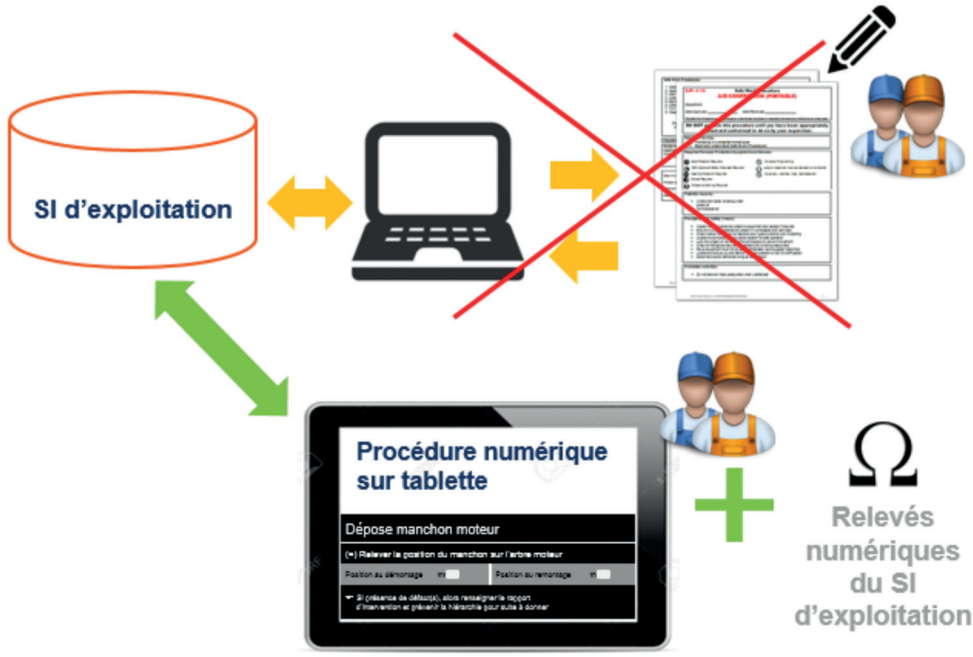


Figure 3 – Dossier de Réalisation des Travaux électroniques. ©EDF

En conclusion

Le parc nucléaire d'EDF en exploitation, constitué de cinquante-huit réacteurs construits par paliers standardisés, dispose d'un patrimoine de données très riche, propice à l'utilisation des nouvelles techniques d'analyse de données permises par l'essor du Big Data.

Après une phase de démonstration de l'intérêt de ces techniques, la généralisation de l'usage des outils de « Data Analytics », menée conjointement aux démarches d'amélioration du « patrimoine données », constitue un levier au service des deux enjeux majeurs du parc nucléaire français : la sûreté et l'amélioration de la performance de l'exploitation et de la maintenance.

Big Data, mutualisation et exclusion en assurance

Par Rémi STEINER
Ingénieur général des Mines, Conseil général de l'Économie

Par essence, les activités d'assurance reposent sur l'exploitation de données statistiques. Il ne fait donc pas de doute que l'explosion quantitative des données numériques et l'amélioration des techniques qui donnent sens à ces données, en d'autres termes le Big Data, constituent des ferments majeurs de transformation. Il s'agit là de l'un des aspects de la mutation numérique à laquelle les assureurs sont confrontés ; le bouleversement des canaux de distribution, la dématérialisation des contrats et l'automatisation des opérations constituent également des enjeux technologiques et commerciaux majeurs.

Une meilleure exploitation des données, à toutes les étapes de la vie d'un contrat d'assurance, est susceptible d'affecter en profondeur le métier d'assureur et de faire émerger de nouveaux terrains de concurrence. Le Big Data pourrait induire une sélection plus pertinente et une plus juste tarification des risques à la souscription d'un contrat ; l'analyse de données issues de véhicules, de capteurs de santé ou d'autres objets connectés pourrait favoriser pendant la durée de vie du contrat des échanges mutuellement profitables à l'assureur et à l'assuré, conduisant à une meilleure prévention des risques et à des services additionnels appréciables ; le Big Data, encore, est susceptible d'améliorer la lutte contre la fraude ou l'identification des bénéficiaires de contrats en déshérence.

Mutualisation ou sélection des risques ?

La mutualisation des risques est au fondement de l'assurance

La mutualisation des risques constitue le fondement des activités d'assurances : un ensemble d'individus, soumis à un même aléa, peuvent trouver un intérêt à mettre leur sort en commun. Ils vont préférer payer une prime d'assurance certaine plutôt que de supporter l'éventualité d'un sinistre d'un montant plus important.

Ils vont implicitement considérer, comme leur assureur commun, que la probabilité d'un sinistre est stable dans le temps, que leur exposition au risque est suffisamment homogène et que les risques individuels sont relativement indépendants. Chaque assuré admet de bonne grâce, une année donnée, de payer sans aucun retour une prime pour un risque qui ne se réalise pas parce que, l'année suivante peut-être, son tour viendra d'être frappé par la matérialisation de ce risque et qu'il sera immunisé par le bénéfice de la mutualisation.

Les assurés de cette communauté lient dans une certaine mesure leurs intérêts : si certains, contre toute attente, s'avéraient exagérément vulnérables, imprudents ou négligents au regard du risque considéré, les primes d'assurance de tous pourraient être solidairement révisées à la hausse.

La concurrence entre assureurs impose la sélection des risques

La solidarité entre assurés pourrait être absolue dans un monde où les activités d'assurances seraient couvertes par un monopole, comme dans le cas de l'assurance maladie obligatoire en France. L'assureur qui bénéficierait de ce monopole aurait la faculté d'accepter tous les clients, quel que soit le niveau de risque qu'ils présentent *a priori* ; il pourrait leur appliquer la même tarification, nonobstant l'hétérogénéité de ce niveau de risque *a priori* ; et la solidarité jouerait à plein.

Mais en situation de libre concurrence, si, face à un aléa donné, une sous-population peut être distinguée et identifiée comme présentant un moindre risque, un nouvel assureur est susceptible d'offrir à cette sous-population, à leur avantage mutuel, des conditions spécifiques plus favorables. Si chaque assuré agit au mieux de son intérêt, ces clients présentant un moindre risque migrent vers le nouvel assureur ; et l'assureur ancien se retrouve avec une population d'assurés en moyenne plus risqués qu'il ne l'avait anticipé. Sauf à accepter de travailler à perte, il est conduit à augmenter ses tarifs.

Aussi les idées de mutualisation et de sélection des risques ne sont-elles pas antagonistes : au contraire, elles se marient nécessairement. En économie de marché, un assureur est conduit à segmenter sa clientèle en fonction de son appréciation des divers niveaux de risque et à appliquer des tarifs différents à des clients différents au regard du risque couvert par le contrat d'assurance. Le droit de refuser purement et simplement une demande d'assurance lui est en outre reconnu.

Les difficultés de la segmentation

Pour une multitude de raisons, la segmentation est un art difficile. À supposer qu'il existe des caractéristiques ou des comportements puissamment discriminants et stables dans le temps, il demeurerait très difficile pour l'assureur de capturer cette information. Le rattachement des demandeurs d'assurance à un segment de risque donné, pour des raisons pratiques, s'appuie habituellement sur des données simplistes (l'âge, le lieu de résidence…), de sorte que les segments ne peuvent être parfaitement homogènes. Même quand un certain paramètre paraît fortement influencer la sinistralité, ce paramètre peut ne pas caractériser le facteur de risque lui-même, mais lui être simplement corrélé : l'âge n'est, par exemple, qu'un révélateur imparfait de l'imprudence et de l'inexpérience au volant.

Il existe habituellement une certaine asymétrie d'information : l'assuré n'a aucune raison de livrer à l'assureur les informations qui l'identifieraient comme présentant un risque *a priori* élevé, justifiant ainsi une prime d'assurance plus importante. Et cette asymétrie d'information peut induire un effet d'anti-sélection : si un assureur soumet à un même échelon tarifaire un ensemble de clients qu'il distingue mal les uns des autres, les moins risqués d'entre eux, conscients de la cherté relative des conditions qui leur sont faites, pourraient préférer s'abstenir de s'assurer, ou s'assurer à de meilleures conditions auprès d'un assureur moins myope. Seuls demeureraient à cet échelon tarifaire les clients les plus exposés au risque, ce que l'assureur, sur la base d'une appréciation globale, a peut-être mal pris en compte dans son tarif. Dans un article fondateur, George A. Akerlof (1970) a montré que de telles asymétries d'information pouvaient empêcher la rencontre de l'offre et de la demande, et donc l'existence même d'un marché.

Les changements permis par le Big Data

Des perspectives nouvelles

L'émergence du Big Data prend des formes variées. D'abord celles de volumes de données et de capacité de traitement informatique qui ouvrent des horizons nouveaux en termes d'analyse statistique. Certes, la profusion des données et la puissance de calcul ne peuvent magiquement résoudre des difficultés de principe liées à la segmentation de la clientèle, ni *a fortiori* valider des tarifs individualisés. Mais il existe des marges de progrès considérables, liées en particulier à l'élargissement de l'accès aux données publiques. Ainsi, les données d'accidentologie automobile ou les données de santé sont aujourd'hui difficilement accessibles et sous-exploitées. Anonymisées autant que nécessaire, elles pourraient être demain plus librement ouvertes à l'analyse et livrer des informations d'un intérêt majeur, non seulement pour les assureurs, mais aussi en faveur d'une meilleure prévention des risques.

Par ailleurs, le Règlement général sur la Protection des Données [1] (RGPD) vient d'établir un nouveau cadre réglementaire d'utilisation des données personnelles. Dans le cadre ancien de la loi du 6 janvier 1978, les entreprises d'assurances, comme toutes les entreprises, étaient soumises à un régime d'autorisation ou de déclaration *ex ante* des traitements informatisés qu'elles mettaient en œuvre. À ce titre, la CNIL encadrait strictement le choix et l'utilisation des données personnelles nécessaires à la passation et à la gestion des contrats d'assurances [2], qu'il s'agisse de l'étude des besoins spécifiques de chaque demandeur ; de l'examen, de l'acceptation, du contrôle et de la surveillance du risque ; de la gestion des contrats de la phase pré-contractuelle à la résiliation du contrat. À compter du 25 mai 2018, le nouveau cadre européen a fait disparaître cet encadrement *a priori* pour le remplacer par un régime plus souple d'autorégulation et de contrôle *a posteriori*. Cette évolution est susceptible de favoriser pour les entreprises d'assurances, sous leur responsabilité et avec des garde-fous importants, une plus grande liberté d'innovation en ce qui concerne le recueil et le traitement de données personnelles.

L'un des terrains privilégiés de développement du Big Data pourrait être lié à l'essor de capteurs et d'objets connectés susceptibles d'objectiver les habitudes de vie d'un assuré, avant et pendant la durée de vie de son contrat d'assurance. Style de conduite automobile, activité sportive, données physiologiques, ou encore localisation sont de plus en plus souvent enregistrés par les instruments de la vie quotidienne : véhicules, smartphones, montre connectée, pèse-personne ou autres appareils ménagers. Dans la mesure où de telles données personnelles seraient objectivement de nature à influencer l'appréciation par l'assureur du niveau de risque d'un contrat, et sous réserve du consentement de l'assuré, elles pourraient demain être intégrées à la politique d'acceptation et de tarification des risques de certains assureurs.

De la segmentation à l'analyse comportementale

De premières expérimentations ont lieu : Generali France propose aux entreprises qui souscrivent à son offre d'assurance complémentaire collective de santé un programme [3] qui vise à inciter leurs salariés à des actions de prévention et à une meilleure hygiène de vie : examens de santé, dépistages, vaccinations, objectifs personnalisés d'activité physique dont l'atteinte est mesurée au moyen d'un bracelet connecté, participation à des événements sportifs, engagement de ne pas fumer, suivi quotidien du diabète… Dans le champ de l'assurance automobile, les conditions de souscription au contrat Allianz Conduite Connectée s'appuient sur un score de conduite apprécié à partir de données collectées par un smartphone pendant une période de test (position GPS, nombre et intensité des accélérations et freinages, prises de virages, date et heure de départ et d'arrivée d'un trajet identifié). En cas de souscription, le véhicule est équipé à demeure d'un boîtier dédié à la collecte de ces informations, qui se substitue au smartphone.

De telles formes d' « assurance comportementale » semblent appelées à se répandre, dans la mesure où elles tendent à assurer une convergence d'intérêt entre l'assureur et l'assuré et à prévenir ainsi l'aléa moral, effet pervers classique de l'assurance : il est courant qu'un individu, dès lors qu'il s'assure contre un risque, infléchisse son comportement et prenne moins de précautions pour se protéger contre la survenance de ce risque. Un mécanisme d'assurance comportementale soigneusement calibré apparaît comme un moyen prometteur de combattre cet effet : à rebours de la situation traditionnelle, la signature d'un contrat d'assurance pourrait conduire un individu à

(1) Règlement 2016/679 du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données.
(2) Délibération n° 2013-212 du 11 juillet 2013 concernant les traitements automatisés de données à caractère personnel relatifs à la passation, la gestion et l'exécution des contrats mis en œuvre par les organismes d'assurances, de capitalisation, de réassurance, d'assistance et par leurs intermédiaires (norme simplifiée n°16).
(3) Generali Vitality.

mieux prévenir un risque contre lequel il s'est pourtant couvert. Si tel devait être le cas, l'intérêt général commanderait d'encourager ces formes de contrat. Elles ont pourtant suscité, en France particulièrement, des commentaires effarouchés, mettant en avant le risque d'une immixtion excessive de l'assureur dans la vie privée des assurés.

De moindres asymétries d'information

Faudrait-il élever des barrières pour éviter qu'un assureur en sache trop sur ses clients ? La question n'est pas nouvelle mais l'émergence du Big Data exacerbe son acuité. N'est-il pas légitime et d'un commun intérêt qu'assureur et assurés puissent s'engager en connaissance de cause, avec l'information la plus complète et la plus objective sur les risques encourus par l'un et l'autre ? Par rapport au monde d'hier, où l'information était rare, l'assureur distant de l'assuré, les politiques de risque opaques et la dissimulation facile, la profusion des informations peut permettre à un assureur et à un assuré, l'un et l'autre de bonne foi, de conclure un contrat dans de meilleures conditions de confiance.

Dans une certaine mesure, parce que toutes les données et les statistiques du monde ne feront jamais disparaître l'incertitude de l'avenir, le Big Data pourrait atténuer le « voile d'ignorance » qui obscurcit la perception des assureurs – et aussi celle des assurés. Les uns et les autres seraient mieux en mesure d'apprécier les risques, et les asymétries d'information pourraient notablement diminuer. Il n'y a aucune raison que les besoins d'assurances disparaissent, mais le marché pourrait être demain plus compétitif et plus transparent.

Big Data et exclusion

La crainte d'un marché plus sélectif

Pour les raisons qu'on a vues, il est probable que ce marché plus compétitif et transparent sera aussi un marché sélectif, où certains demandeurs seront exclus, partiellement ou totalement, du bénéfice d'un contrat d'assurance, ou bien soumis à un tarif élevé. En réalité, cette situation recouvre deux cas de figure distincts, qui soulèvent des questions de principe totalement différentes.

Le premier cas de figure est celui caractérisé, soit par l'aléa moral, soit par la trop faible aversion au risque de certains assurés : l'assuré qui commet des infractions au code de la route, qui ne prend pas de précautions contre le cambriolage et qui ne se préoccupe nullement de sa santé, pour des raisons propres à sa personnalité et à son comportement, et sans que cet aspect soit toujours détectable par l'assureur, présente un risque plus élevé que la moyenne. À l'extrême, certains assurés peuvent être décrits comme des passagers clandestins qui tirent un profit indu de leur couverture assurantielle, au détriment à la fois de l'assureur et des autres assurés. On peut s'interroger sur les conditions de mutualisation de ce type de risque et estimer que sa meilleure identification est conforme à l'intérêt général.

Tout autre est la situation où un risque aggravé est une donnée intangible, une caractéristique propre au candidat à l'assurance et sur laquelle il n'a pas de prise : on pense ici notamment à sa situation de santé ou à son exposition aux catastrophes naturelles. Certaines pratiques d'exclusion peuvent apparaître choquantes et appeler soit une renonciation spontanée des assureurs à recourir à certains critères de risque, soit une réglementation *ad hoc*. La crainte par l'assureur d'une atteinte à sa réputation ou l'édiction de règles d'ordre public aboutissent toutes deux à une mutualisation du risque, c'est-à-dire à l'acceptation plus ou moins tacite par les assurés d'un renchérissement de leur contrat en contrepartie d'une absence de discrimination par rapport au critère de risque considéré.

L'intervention de la réglementation

Il existe de nombreux exemples de réglementations qui imposent aux assureurs – et aux assurés – une telle mutualisation forcée des risques. La couverture du risque de catastrophe naturelle, que la loi [4] associe obligatoirement à tout contrat d'assurance «multirisques habitation», en est un.

Autre cas, la loi Évin [5] a disposé que lorsque des salariés sont garantis collectivement par un contrat complémentaire santé, l'organisme ne peut refuser de maintenir le remboursement ou l'indemnisation des frais occasionnés par une maladie aux personnes affiliées au contrat tant que celles-ci le souhaitent, et ne peut non plus réduire les garanties souscrites, aux conditions tarifaires de la catégorie dont les assurés relèvent. L'organisme ne peut ultérieurement augmenter le tarif d'un assuré ou d'un adhérent en se fondant sur l'évolution de l'état de santé de celui-ci. Si l'organisme veut majorer les tarifs d'un type de garantie ou de contrat, la hausse doit être uniforme pour l'ensemble des assurés ou adhérents souscrivant ce type de garantie ou de contrat.

Un exemple plus discuté tient à la mutualisation des risques selon le sexe, alors même que des écarts statistiques sont avérés. La directive 2004/113/CE du 13 décembre 2004, mettant en œuvre le principe de l'égalité de traitement entre les femmes et les hommes dans l'accès à des biens et services et la fourniture de biens et services, avait accordé aux États membres la faculté de déroger à une application stricte du principe d'égalité de traitement entre les hommes et les femmes en matière d'assurance. Mais un arrêt de la Cour de Justice de l'Union européenne [6] du 1er mars 2011 a jugé que l'application de primes différentes aux hommes et aux femmes constituait une discrimination fondée sur le sexe, prohibée par la Charte des droits fondamentaux de l'Union européenne.

Big Data et risque aggravé de santé

Un enjeu de société absolument prééminent est celui de l'accès des personnes présentant un risque aggravé de santé à l'assurance décès, et donc de leur accès à l'emprunt immobilier et à la propriété. Le Big Data est-il susceptible, en dévoilant avec impudeur l'espérance de vie de chacun, de frapper d'ostracisme une partie de la population ?

S'il est légitime de se préoccuper des évolutions à venir, force est de reconnaître que l'exclusion est déjà une réalité. Les statistiques arrêtées au titre de l'année 2016 dans le cadre de la convention AERAS (s'Assurer et Emprunter avec un Risque aggravé de Santé) montrent que 514 449 demandes d'assurance décès sur 3 447 038 (soit 15 %) ont été identifiées comme présentant un risque aggravé de santé. Plus de deux fois sur trois, après examens complémentaires, une proposition d'assurance aux conditions normales du contrat a été faite.

Mais 133 025 demandes (soit 4 % du total) ont donné lieu à une proposition comportant une surprime ; dans la moitié des cas, cette surprime était supérieure à 50 % et elle excédait même 300 % dans 2 117 cas. Des refus purs et simples ont été prononcés dans 18 046 cas ; une proposition sans surprime mais avec exclusion ou limitation de garantie était faite dans 2 569 cas. Pour des raisons indéterminées, 89 664 demandes d'assurance sont en outre restées inabouties du fait des assurés. Alors qu'on peut de surcroît présumer une part inquantifiable d'autocensure en amont des demandes, l'exclusion en matière d'assurance est loin d'être un phénomène marginal.

Sous la pression des associations de patients et des pouvoirs publics, l'article 190 de la loi du 26 janvier 2016 de modernisation de notre système de santé a permis de consacrer le principe d'un « droit à l'oubli », en faveur notamment des personnes ayant été affectées d'un cancer. Ce droit,

(4) Article L. 125-1 du Code des assurances.
(5) Loi n° 89-1009 du 31 décembre 1989 renforçant les garanties offertes aux personnes assurées contre certains risques, article 2.
(6) Arrêt de la CJUE du 1er mars 2011, Association belge des Consommateurs Test-Achats ASBL.

inscrit aux articles L. 1141-5 et L. 1141-6 du Code de la santé publique, doit être progressivement étendu aux pathologies autres que cancéreuses, notamment les pathologies chroniques, dès lors que les progrès thérapeutiques et les données de la science attestent de la capacité des traitements concernés à circonscrire significativement et durablement leurs effets [7].

Il existe donc d'immenses voies de progrès en faveur d'une meilleure inclusion et il est tout à fait possible que les effets du Big Data, par le jeu combiné d'une meilleure exploitation statistique des données publiques et privées, d'une meilleure transparence du marché de l'assurance et d'un renforcement de la concurrence, se traduisent en définitive par une moindre pusillanimité des assureurs et par une diminution des phénomènes d'exclusion. Tel est en tout cas l'objectif qui devrait continuer à guider les pouvoirs publics.

Bibliographie

AERAS, « s'Assurer et Emprunter avec un Risque aggravé de Santé », Convention du 6 juillet 2006, avenants successifs, grille de référence et statistiques annuelles.

AKERLOF G. (1970), "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism", *The Quarterly Journal of Economics*, Vol. 84, n° 3 (Aug.), pp. 488-500.

BERBAIN C. & SALAMANCA E. (2015), *L'assurance de demain, Reconnaîtrons-nous notre assureur en 2030 ?*, Mémoire de troisième année du Corps des Mines, juillet.

CNIL (2014), *Pack de conformité assurance*.

CHARPENTIER A. & BARRY A. (2015), « Passer d'une analyse de corrélation à une interprétation causale », *Risques*, n° 99.

CHARPENTIER A., DENUIT M. & ÉLIE R. (2015), « Segmentation et Mutualisation, les deux faces d'une même pièce ? », *Risques*, n° 103.

DUMORA R. (2018), « Les nouveaux chemins de l'assurance », in *Les Banques face à leur avenir proche*, Eyrolles, pp. 265-288.

FÉDÉRATION FRANÇAISE DE L'ASSURANCE (2017), *Convention AERAS : statistiques 2016*.

FROMENTEAU M., RUOL V. & ESLOUS L. (2011), « Sélection des risques : où en est-on ? », in *Les Tribunes de la Santé* 2011/2, n° 31.

VILLANI C. (2018), « Donner un sens à l'intelligence artificielle, Focus 2 – La santé à l'heure de l'IA », Rapport au Premier ministre.

(7) La grille de référence annexée à la Convention AERAS a été mise à jour pour la dernière fois le 30 mars 2017.

Le Big Data en agriculture

Par Véronique BELLON-MAUREL
ITAP, Irstea, Montpellier Supagro, Université de Montpellier
Pascal NEVEU
MISTEA, INRA, Montpellier Supagro, Université de Montpellier
Alexandre TERMIER
Université de Rennes, Inria, CNRS, IRISA
et Frédérick GARCIA (1)
MIAT, UR875, Université de Toulouse, INRA

Les travaux sur le Big Data agricole sont tous très récents, et encore peu nombreux puisqu'on n'en dénombre que quelques dizaines (Kamilaris *et al.*, 2017 ; Wolfert *et al.*, 2017). Cependant, même si l'adoption est plus lente que dans d'autres secteurs, il n'en demeure pas moins que le potentiel du Big Data en agriculture est énorme, tout autant que les enjeux et les challenges à relever.

Les enjeux du Big Data en agriculture

L'agriculture vit des transformations majeures qui lui imposent de s'adapter. Le changement climatique et la dégradation des sols font rapidement évoluer les conditions pédoclimatiques et l'agriculteur ne maîtrise plus son outil de production. Les pouvoirs publics et les consommateurs ont des exigences nouvelles par rapport à la production agricole : agro-écologie, bien-être animal, arrêt du glyphosate… Tout ceci impose un changement de modèle agricole, qui doit être rapide mais qui est d'autant plus difficile que les agriculteurs ne connaissent pas forcément les pratiques à mettre en œuvre, et que les référentiels n'existent pas encore, par manque de recherche. Il est donc indispensable de construire rapidement une connaissance pour accompagner l'agriculteur dans un processus de production toujours plus complexe, combinant recherche de compétitivité, respect des réglementations, recherche d'une meilleure valorisation financière et évolution des pratiques. Nous faisons l'hypothèse que les données en lien avec la production agricole sont sur une trajectoire de massification qui permettra de constituer des méga-données – ou Big Data – et peuvent être utilisées dans des analyses de Big Data pour fournir à l'agriculteur des outils d'aide à la décision.

Les enjeux du Big Data agricole sont de plusieurs ordres : ils touchent la production agricole et la commercialisation, mais également les pouvoirs publics pour la gestion de l'espace et la mise en place de politiques publiques. Concernant la production agricole, l'analyse des données massives est utile à la fois dans les choix stratégiques de l'exploitation (quelles espèces ? quelles rotations ?) et dans les choix tactiques (interventions dans l'itinéraire technique). Des modèles de fonctionnement plus précis, car intégrant mieux les conditions locales, sont déduits de l'analyse, en particulier pour la gestion des intrants ; les semenciers parlent de « semis prescriptif », à savoir de conseils sur la densité des semis et les variétés à semer, un service qu'ils envisagent de vendre aux agriculteurs. De la même manière, pour les animaux, on pourra mieux connaître la manière dont chaque individu s'alimente ou supporte des stress, et ainsi ajuster son alimentation ou les interventions. Ainsi, dans ces domaines de l'agriculture et de l'élevage de précision, des modèles de fonctionnement plus

(1) Les auteurs appartiennent à l'Institut Convergences Agriculture Numérique #DigitAg, qui bénéficie d'une aide de l'État gérée par l'Agence nationale de la Recherche au titre du programme d'Investissements d'avenir portant la référence ANR-16-CONV-0004.

précis peuvent être inférés de l'analyse du Big Data. L'aspect prédictif des connaissances issues du Big Data est aussi un enjeu. Par exemple, dans certains domaines dans lesquels la connaissance est limitée, comme pour certaines approches de l'agro-écologie ou de l'agroforesterie, il s'agit de mettre en place des expériences, y compris chez les agriculteurs, dans des conditions pédoclimatiques très variées, pour en extraire les règles agronomiques de fonctionnement de ces nouveaux écosystèmes et pouvoir conseiller les agriculteurs sur ces nouveaux itinéraires culturaux. Le numérique facilite cette approche de *crowdsourcing* et de sciences participatives (Minet *et al*, 2017), ce qui permettra aussi une meilleure diffusion des innovations agricoles. Du côté des pouvoirs publics, l'analyse des données massives permet d'estimer les rendements des cultures, information précieuse pour garantir la sécurité alimentaire, de suivre les épisodes épidémiques et de prendre des mesures sanitaires. Enfin, dans le cadre des services d'assurance indicielle, qui s'avèrent être bénéfiques aux petits agriculteurs, le Big Data est le socle indispensable pour construire des indices effectifs apportant une réelle protection aux agriculteurs ; cela justifie pleinement l'augmentation des investissements publics dans la collecte des données qui peuvent faciliter l'expansion des marchés de l'assurance indicielle (Castillo *et al.,* 2016).

Les enjeux étant posés, intéressons-nous aux spécificités du Big Data agricole.

Le Big Data agricole, de la donnée à l'usage

Les sources de données du Big Data agricole

Les sources de données du Big Data agricole sont multiples et prolifèrent. Les premières sont les images satellitaires, dispositifs qui datent de plus de quarante ans, mais qui connaissent aujourd'hui un déploiement inédit avec le lancement de nouveaux satellites. Par exemple, la constellation Sentinel délivre des images gratuitement à une très haute fréquence temporelle (tous les cinq jours), ce qui offre des opportunités totalement nouvelles à la recherche et aux entreprises, en permettant par exemple de fonder un conseil agronomique. À cela s'ajoutent des dispositifs de mutualisation comme le dispositif GeoSud THEIA (https://www.theia-land.fr/fr/projets/geosud), qui permet l'accès à moindre coût à des images satellitaires plus sophistiquées, multispectrales et susceptibles d'être programmées. Un autre facteur de prolifération des données est la multiplication des objets connectés en agriculture (Tzounis *et al.*, 2017) : stations météorologiques, pièges à insectes, capteurs d'humidité du sol et compteurs d'eau connectés en irrigation, mais aussi divers capteurs installés sur les animaux pour évaluer leur état (état de santé, présence de chaleurs…), sur les robots de traite (quantité et qualité du lait) ou sur les automates d'alimentation. Les agro-équipements sont aussi de plus en plus fréquemment équipés de capteurs, pour l'agriculture de précision (nécessité de connaître les besoins de la plante pour lui apporter exactement ce dont elle a besoin), mais aussi pour la maintenance prédictive. Les exigences de traçabilité des productions agricoles se traduisent aujourd'hui par des systèmes de lecture automatisée, avec les puces RFID, ou par la saisie manuelle des interventions agricoles à partir de smartphones avec transmission directe vers les applications des éditeurs de logiciels, ce qui remplace les cahiers de culture en papier. Le défi est d'automatiser l'acquisition de données afin qu'elle n'ait virtuellement aucun coût (Wolfert *et al.*, 2017). Enfin, les méthodes de phénotypage rapide, incontournables pour raccourcir le cycle de production de nouvelles semences, sont également sources de données massives à mettre en relation avec les données génotypiques (Halewood *et al.*, 2018). La section suivante montre en quoi ces données massives peuvent être qualifiées de Big Data.

Les 6 « V » du Big Data agricole

Dans le domaine des sciences de données, le Big Data se caractérise par les «V» et le Big Data de l'agriculture n'échappe pas à cette règle mais avec certaines spécificités.

- Le «Volume» toujours croissant des données est lié au développement des technologies, à l'usage du web et, au-delà de la captation de la donnée, à la réduction des coûts de stockage de l'information qui permet de la réutiliser.

- La «Variété» est la caractéristique forte du Big Data agricole. Au-delà des différentes natures des informations (texte, mesure, image, spectre, simulation...), le domaine de l'agriculture exige de prendre en compte des sources de données nombreuses et extrêmement diverses, provenant :

  - des échelles allant du territoire à la génomique en passant par différents niveaux tels que l'exploitation, la population ou l'individu (plante ou animal) ;
  - des informations spatiales et cartographiques pouvant décrire des zones, des trajectoires, des opérations culturales ou des sites hétérogènes d'acquisition ; des interactions et échanges qui se produisent à toutes ces échelles avec l'environnement (sol, eau, climat, bioagresseurs…) ;
  - des évolutions (stades de croissance, modes de conduite) ;
  - des transformations qui vont de la ferme à l'assiette ;
  - des impacts sociaux, économiques, environnementaux, sanitaires.

Au-delà des données, c'est la variété des acteurs, qui les produisent et qui les manipulent, qui est en soi un défi en agriculture. Ces acteurs (agronomes, généticiens, biologistes, statisticiens, industriels, distributeurs ...) ont par essence des vocabulaires, des sémantiques et des modes de production de données très hétérogènes. Cette forte variété rend difficile l'intégration des données. Cela demande donc de nouvelles approches pour gérer et analyser les données en prenant en compte les dimensions spatiales, temporelles mais aussi organisationnelles (filières, circuits courts, normes sanitaires).

- La «Vélocité» dans l'agriculture correspond à de nouvelles générations d'outils pour une aide à la décision où la vitesse d'acquisition des données et la vitesse pour fournir un résultat aux utilisateurs sont cruciales. Par exemple, ces outils peuvent être liés à l'exploitation d'un grand nombre de capteurs équipant des animaux, à l'utilisation du *crowdsourcing*, aux plateformes de phénotypage à haut-débit, ou encore à l'agriculture de précision. Un des enjeux est de pouvoir gérer et analyser de grandes masses de données en temps réel.

- D'autres «V» peuvent être mentionnés. La «Véracité» et la «Validité» qui caractérisent la qualité de sources de données ; la «Visualisation» qui est une problématique pour de grands ensembles de données complexes ; la «Visibilité» de faits ou d'éléments pertinents qu'il faut extraire de gros volumes de données ; enfin l'analyse de toutes ces données peut poser des problèmes de respect de la «Vie privée» d'exploitants agricoles.

Enfin, le véritable enjeu du Big Data, c'est la Valeur qu'on peut en obtenir. Cet enjeu dépend de notre capacité à structurer le Big Data. Ceci permettra de réutiliser et d'agréger des données de différentes façons mais aussi de pouvoir les lier avec d'autres données. Autrement dit, l'innovation autour du Big Data agricole sera indispensable pour créer de la valeur et cela passe par la structuration des données et le développement de nouvelles approches d'analyse.

L'analyse du Big Data agricole

L'analyse peut avoir pour but de fournir aux différents utilisateurs, comme les agriculteurs ou les conseillers agricoles, une aide pour prendre la bonne décision au bon moment, par des indicateurs, des alertes ou des explications. Elle peut aussi avoir pour but de découvrir de nouvelles connaissances, aussi bien générales, par la masse des données collectées, que localisées sur une exploitation, une parcelle ou un animal particulier grâce à un suivi fin. Un domaine historique est l'analyse d'images, que ce soient des images satellitaires (ex : comprendre l'utilisation des sols et la dynamique des cultures) ou des images prises depuis un aéronef ou un drone (ex : suivre l'évolution de la croissance des feuilles et les éventuelles maladies). Des approches d'apprentissage automatique

ou de fouille de données peuvent être appliquées pour corréler les grandeurs mesurées avec des variables d'intérêt ou effectuer des prédictions (Kamilaris *et al.*, 2017). Par exemple, dans les exploitations laitières, il est désormais possible de suivre la température d'un animal et son activité au long de la journée : un enjeu est d'analyser ces données pour prédire au plus tôt les périodes de fertilité, ou les maladies (Steeneveld *et al.*, 2015).

Une difficulté est l'adaptation des méthodes de fouille de données aux caractéristiques multi-scalaires (temporelles ou spatiales) des données agricoles. D'autres difficultés sont la gestion de l'incertitude dans les données et la réduction du nombre de paramètres dans les méthodes utilisées afin de simplifier leur utilisation.

Les acteurs du Big Data agricole

Aujourd'hui de nombreux acteurs se positionnent : les agro-équipementiers avec leurs outils connectés, l'agrofourniture qui ambitionne de vendre des services plutôt que des intrants, les grosses coopératives qui génèrent de la donnée, mais aussi des acteurs exogènes au monde agricole comme des financiers (capital-risqueurs) et des spécialistes de la gestion de données par les technologies *Cloud* (Google, IBM, Fujitsu…) (Wolfert *et al.,* 2017). Par exemple Google Venture a investi 15 millions de dollars dans une société capable de mettre en œuvre du Big Data pour aider les exploitants à mieux vendre leurs produits [(2)]. De nouveaux jeux d'acteurs et de nouvelles relations de pouvoir se mettent en place dans la chaîne de valeur de l'agro-alimentaire, au travers de la maîtrise du Big Data. Les pouvoirs publics doivent jouer un rôle dans cet espace : mettre à disposition des infrastructures, favoriser l'égalité des territoires par la connectivité, veiller à éviter les situations de monopoles…

Conclusion

Malgré l'importance des enjeux en termes de développement et viabilité économiques, l'adoption des innovations liées au numérique et au Big Data par les acteurs du monde agricole reste freinée par de nombreuses barrières (Kamilaris *et al.*, 2017), dont les principales sont techniques : manque d'expertise et de ressources humaines dans le domaine des technologies de l'information appliquées à l'agriculture, offre limitée et caractère difficilement accessible des infrastructures de données offrant à la fois des services de stockage de l'information, mais également d'analyse des Big Data, et, lorsque ces offres existent, difficulté de coupler entre elles ces différentes sources de données, du fait de leur grande hétérogénéité (le « V » de Variété) et de l'absence d'ontologies largement partagées. Par ailleurs, le développement du Big Data en agriculture est également freiné par l'absence d'un modèle économique clair et attractif pour l'ensemble de la chaîne de la donnée, intégrant les agriculteurs, les transformateurs, les distributeurs, ainsi que les consommateurs, les citoyens et les services publics également dans un rôle de producteurs de données. Il s'agit d'inventer ici un fonctionnement économique en accord avec les principes actuellement mis en place en France et Europe autour du droit, de la propriété intellectuelle et de l'ouverture des données et, plus globalement, des produits numériques qui autorise un partage équitable de la plus-value créée par le Big Data et qui favorise la création et le développement d'innovations numériques en termes d'outils et d'usages pour le domaine agricole.

Au-delà de ces obstacles, il est également nécessaire de prévenir certains risques associés au développement du numérique en agriculture, comme la perte potentielle d'autonomie des agriculteurs avec la possible hyper-technologisation du domaine, et l'hypercentralisation de la chaîne de la donnée par quelques acteurs globaux, ou encore l'augmentation de la fracture numérique entre les agricultures des pays développés et en développement.

(2) https://venturebeat.com/2015/05/19/google-ventures-leads-15m-investment-in-big-data-for-farmers/

Tous ces sujets, techniques et sociaux, font aujourd'hui l'objet de nombreuses recherches, en particulier au sein de l'Institut Convergences Agriculture Numérique #DigitAg.

Références

CASTILLO M.J., BOUCHER S. and CARTER M. (2016), "Index Insurance: Using Public Data to Benefit Small – Scale Agriculture", *International Food and Agribusiness Management Review*, Special Issue – Vol.19 Issue A, 93-114.

COOPER J., NOON M., JONES C., KAHN E. and ARBUCKLE P. (2013) "Big Data in Life Cycle Assessment", *J Ind. Ecology* 17 (6), 796-799.

HALEWOOD M., CHIURUGWI T., SACKVILLE HAMILTON R., KURTZ B., MARDEN E., WELCH E., MICHIELS F., MOZAFARI J., SABRAN M., PATRON N., KERSEY P., BASTOW R., SHAWN DORIUS S., DIAS S., McCOUCH S. and POWELL W. (2017), "Plant genetic resources for food and agriculture: opportunities and challenges emerging from the science and information technology revolution", *New Phytologist* 217, 1407-1419.

KAMILARIS A., KARTAKOULLIS A. and PRENAFETA-BOLDU F.X. (2017), "A review on the practices of big data analysis in agriculture", *Computers and Electronics in Agriculture.* 143, 23-37.

MINET J., CURNEL Y., GOBIN A., GOFFART J.P., MÉLARD F., TYCHON B., WELLENS J. and DEFOURNY P. (2017), "Crowdsourcing for agricultural applications: A review of uses and opportunities for a farmsourcing approach", *Computers and Electronics in Agriculture* 142, 126-138.

STEENEVELD W. and HOGEVEEN H. (2015), "Characterization of Dutch dairy farms using sensor systems for cow management", *Journal of Dairy Science*, 709-717.

TZOUNIS A., KATSOULAS N., BARTZANAS, T. and KITTAS C. (2017), "Internet of things in agriculture, recent advances and future challenges", *Biosystems Engineering,* 164, 31-48.

WOLFERT S., GE L., VEROUW C. and BOGAARDT M.J. (2017), "Big Data in smart farming – A review", *Agricultural Systems*, 153, 69-80.

Les Big Data : quelles perspectives pour la statistique publique ?

Par Didier BLANCHET
Insee, directeur des Études et synthèses économiques
et Pauline GIVORD
Insee, responsable du SSPLab

L'arrivée des Big Data va-t-elle modifier radicalement la façon de produire les chiffres qui alimentent le débat public et la conduite des politiques économiques et sociales ? L'essentiel de cette production repose actuellement sur les instituts nationaux de statistique. Ils s'appuient sur des sources déjà très diverses et souvent volumineuses : répertoires, recensements, enquêtes, sources administratives, principalement les sources sociales et fiscales. Les enquêtes et les recensements suivent des protocoles aussi stables que possible, qui garantissent la cohérence de leurs résultats dans le temps. L'avantage des sources administratives est de limiter la charge de réponse pour les enquêtés mais leur contenu n'est pas directement formaté pour les besoins de la statistique : leur exploitation nécessite donc d'importants travaux de retraitement. Collecter, exploiter et synthétiser l'ensemble de ces sources constitue le cœur du métier de statisticien public. Son travail est encadré par des accords ou règlements internationaux et soumis, en Europe, à des procédures de revue par les pairs : chaque institut national est régulièrement inspecté par des représentants d'autres instituts, ce qui vise à garantir à la fois la qualité et l'indépendance de la production statistique.

Avoir rappelé ces caractéristiques de la production statistique actuelle permet de mieux cerner les questions que lui pose l'exploitation des Big Data [(1)]. Par Big Data, on entendra principalement les masses de données générées par le développement du numérique : informations directement disponibles sur le web, traces qui y sont laissées par le comportement des internautes, données de transaction, mais aussi les enregistrements issus des réseaux ou des capteurs, comme les données produites par la téléphonie mobile ou encore les données satellitaires… Leur volume et leur caractère souvent peu structuré soulèvent des problèmes de traitement spécifiques. En extraire une information pertinente demande des investissements conséquents qui peuvent se révéler rapidement obsolètes, car l'usage de ces outils numériques est extrêmement évolutif.

Par ailleurs, une bonne partie de ces Big Data est issue de l'activité du secteur privé et détenue par les entreprises qui les génèrent. Faut-il dès lors s'attendre à une marginalisation de la statistique publique avec une reconfiguration radicale du mode de production et de diffusion de l'information économique et sociale ? On devine les risques auxquels on se retrouverait exposé : prolifération d'informations concurrentes sans harmonisation ni stabilité temporelle, neutralité non garantie dès lors que les informations seraient traitées et diffusées sans les procédures de surveillance qui encadrent la statistique publique. La bonne approche pour la statistique publique est plutôt d'explorer les complémentarités de ces nouvelles sources avec celles déjà existantes : comment peuvent-elles se combiner aux sources traditionnelles, et comment les instituts nationaux peuvent-ils progressivement les intégrer à leurs processus de production ?

(1) Le lecteur est renvoyé à BLANCHET & GIVORD (2017) pour une présentation plus complète.

Un premier exemple : la mesure des prix

La mesure des prix permet d'illustrer plusieurs de ces questions. Actuellement, l'essentiel du suivi des prix se fait par collecte directe sur les lieux de vente. Ce mode de recueil a l'avantage d'être applicable à tous les types de biens mais il est lourd et coûteux. Pour mesurer l'indice des prix à la consommation, les enquêteurs de l'Insee relèvent environ 200 000 prix chaque mois dans près de 30 000 points de vente.

L'explosion du numérique offre deux nouveaux modes de recueil. Le premier est de recueillir en temps réel les prix en ligne sur les sites Internet des distributeurs. Ce *webscraping* est mis en œuvre par un projet international conduit hors du champ de la statistique officielle, le *Billion prices project* (BPP). Son origine est un cas de contestation de la statistique officielle, la mesure de l'inflation en Argentine à la fin des années 2000. La défiance vis-à-vis de la mesure des prix est un phénomène classique. Dans le cas argentin, elle s'était trouvée confirmée par des évaluations issues d'autorités locales indépendantes et par des travaux d'économistes : une inflation officielle de l'ordre de 7 % par an et des estimations alternatives de l'ordre de 20 %. Le recours au *scraping* des sites de grandes enseignes avait permis de confirmer cet écart, prouvant du même coup la faisabilité de ce mode de collecte. Le BPP est directement issu de cette expérience. Il a été lancé en 2008 en tant que projet académique, avec l'objectif de couvrir le plus grand nombre possible de pays. La cible symbolique du milliard de prix qui avait donné son nom au projet a été atteinte, en flux annuel, dès 2010. Le changement d'échelle a nécessité la recherche de financements et a conduit à la création d'une entreprise dédiée (www.pricestats.com) qui suit actuellement 15 millions de produits pour 900 détaillants de 50 pays (Cavallo et Rigobon, 2016).

L'extension du projet initial à d'autres pays a montré de manière rassurante que la défaillance observée dans le cas argentin est l'exception plutôt que la règle : ainsi, pour les États-Unis et la zone euro, indices BPP et indices officiels s'avèrent très concordants, notamment sur le très faible niveau de l'inflation des dernières années. Ce résultat est plutôt confortant pour les collectes traditionnelles, mais il pourrait aussi plaider pour leur remplacement progressif par cette nouvelle technique. Ce n'est pas cette voie qui est privilégiée par la majorité des instituts nationaux de statistique. Le *scraping* est certes à l'étude dans certains instituts et, en France, certains prix, tels que ceux des transports aériens ou maritimes, sont d'ores et déjà récupérés sur le web. Mais, pour les biens, la préférence va en général à un autre type de données numériques massives, les données de caisse, issues des facturettes émises lors du règlement des achats en magasin. Elles ont l'avantage d'informer à la fois sur les prix et les quantités achetées, fournissant donc directement les deux types d'informations requises pour la construction de l'indice des prix.

En France, la mobilisation de ces données de caisse a fait l'objet d'un projet entrepris par l'Insee en 2015, d'abord expérimental et qui devrait aboutir en vraie grandeur d'ici 2020. Il bénéficie désormais d'un cadre législatif sécurisé, la loi pour une République numérique ayant prévu les conditions de mise à disposition de ce type de données par les principales enseignes de la grande distribution. La statistique publique mobilisera ainsi les données du secteur privé d'une autre façon que ne le fait le BPP, dans le cadre de relations contractuelles stables avec ces grandes enseignes. L'évolution que représente ce passage aux données de caisse reproduirait à quelques décennies d'écart ce qui s'est passé dans le domaine des sources administratives : obtenir l'accès à ces sources n'a pas non plus été un processus immédiat, certains instituts étrangers continuent d'ailleurs de moins y recourir que ne le fait actuellement la France.

Big Data et *nowcasting* : l'illusion de la vélocité ?

L'exemple des prix illustre aussi plusieurs des différents « V » souvent utilisés pour définir les Big Data. Tout d'abord la volumétrie puisqu'on attend beaucoup du fait de bénéficier de relevés à niveau bien plus fin qu'il n'est possible avec les relevés manuels. Le V de « variété » concerne plutôt la technique du *webscraping*. La variété est dans ce cas d'espèce une contrainte plutôt qu'un atout : c'est le tour de force du BPP de réussir à produire une statistique apparemment pertinente à partir de l'information très disparate recueillie sur les sites des enseignes. De ce point de vue, l'avantage des données de caisse est de se présenter sous un format qui est plus proche de celui des sources traditionnelles, même s'il ne faut pas sous-estimer le coût de la mise sous un format unique de données issues de plusieurs systèmes informatiques.

Le V de « véracité » peut aussi être évoqué, mais sans constituer un avantage comparatif : les prix collectés en rayon, sur le web ou sur les données de caisse, sont tout aussi « vrais » les uns que les autres, avec pour les données de caisse le seul avantage supplémentaire de pouvoir intégrer les rabais offerts lors des achats en magasin. À l'inverse, les données « scrapées » ont le désavantage de ne couvrir ni tous les biens ni tous les modes de vente : elles se limitent par nature aux biens vendus en ligne.

Qu'en est-il du V de « vélocité » ? Il est mis en avant par les promoteurs du BPP et, de fait, le recueil des prix en ligne doit permettre de capter des accélérations ou décélérations rapides des prix en temps quasi réel. Mais ce gain ne serait que de quelques jours par rapport aux indices traditionnels : en France, une première estimation de l'indice des prix à la consommation d'un mois donné est désormais disponible dès la fin de ce mois, et en général très peu révisée lors de sa publication définitive au milieu du mois suivant.

Le gain en vélocité peut-il être plus décisif sur d'autres segments du diagnostic conjoncturel ? En l'état, le diagnostic macro-économique s'appuie sur un enchaînement de sources de finesse croissante. Les enquêtes de conjoncture qualitatives et les indices quantitatifs de production ou de chiffres d'affaires sont publiés mensuellement et constituent les sources principales du diagnostic à court terme, avant la mobilisation des sources administratives et des enquêtes lourdes pour l'élaboration beaucoup plus progressive des comptes annuels détaillés. Depuis le début de 2016, la première estimation des principaux agrégats du trimestre est publiée un mois après la fin de ce trimestre. Il s'agit donc de délais déjà très courts, avec des chiffres certes révisables, et fatalement très révisés, mais qui ont l'avantage de s'appuyer sur des protocoles stables et un recueil représentatif de l'ensemble de l'économie.

Peut-on faire mieux en s'appuyant sur les informations extraites des Big Data ? Plusieurs expérimentations tentent de tirer parti des comportements des internautes, tels que la fréquence des recherches sur certains types de mots-clés sur Google ou la tonalité des échanges sur les réseaux sociaux, pour une capture presque en temps réel d'un certain nombre de phénomènes sociaux ou économiques. L'hypothèse est que ces comportements de recherche sont prédictifs des grandeurs qu'on cherche à évaluer : par exemple, la fréquence des recherches sur les termes d'emploi ou d'assurance chômage est probablement corrélée avec la conjoncture du marché du travail, la recherche d'information sur certains types de produits de consommation et de services avec les achats qui vont se réaliser.

La démarche rappelle évidemment la façon dont l'exploitation des données du web a été mise en avant pour la prévision des comportements électoraux au cours de la période récente, avec la même idée que les mouvements observés sur le web pouvaient prédire les votes de façon plus fiable que les sondages d'opinion traditionnels. Or on sait que les résultats ont été très ambivalents : cette démarche a parfois fait bien mieux que les méthodes classiques, mais elle a fait parfois beaucoup

moins bien, avec le risque que succès comme échecs n'aient été que le fruit des circonstances. La même mésaventure avait affecté un autre type de tentative promue par Google : l'utilisation des comportements de recherche sur certains termes médicaux pour le suivi en temps réel des épidémies de grippe aux États-Unis. Cette expérience n'a fonctionné qu'un temps et a dû être ensuite abandonnée (Lazer *et al.*, 2013). Dans le domaine économique, les travaux montrent en général que ces techniques n'apportent au mieux qu'une information marginale par rapport à celle déjà contenue dans les enquêtes de conjoncture (Bortoli et Combes, 2014). L'apport de ce type de données n'est vraiment substantiel que dans les pays qui ne disposent pas déjà d'un appareil de suivi conjoncturel bien développé.

Des statistiques expérimentales pour combler les *data gaps*

De fait, hormis le cas des données de caisse, c'est pour compléter des champs encore peu couverts par la statistique publique que l'exploitation de ces nouvelles sources semble avoir le plus de potentiel, plutôt que pour remplacer les collectes existantes. L'un de ces domaines est la mesure de l'économie numérique et des nouvelles activités qui en découlent. Des enquêtes européennes harmonisées renseignent certes déjà sur le recours des entreprises et des ménages aux outils numériques. Mais l'exploitation des contenus des sites Internet des entreprises, ou de ce qui se dit de ces entreprises dans la presse en ligne et les réseaux sociaux, pourrait permettre d'en savoir plus sur leur insertion dans l'économie numérique : des expériences de ce type ont été menées au Royaume-Uni ou aux Pays-Bas dans le cadre de partenariats entre acteurs privés et équipes académiques ou institut national de statistique (Nathan et Rosso, 2013 ; Oostrom *et al.*, 2016). Les données bancaires ouvrent aussi des perspectives sur les sources de revenu associées aux nouvelles formes d'emploi en partie issues de la numérisation (Farrell et Grieg, 2016). L'analyse des comportements des consommateurs de services en ligne peut aussi aider à chiffrer la valeur monétaire des services qu'ils en retirent (Cohen *et al.*, 2016 ; Brynjolfsson *et al.*, 2017).

Les explorations portent également sur d'autres champs que la mesure de l'économie *stricto sensu*, en particulier pour évaluer la distance aux objectifs de développement durable (ODD) définis par l'ONU en septembre 2015 dans le cadre de l'Agenda 2030. L'utilisation de données satellitaires, qui permettent une description fine de l'utilisation des sols (artificialisation, type d'agriculture, zones humides…), est envisagée pour définir plusieurs indicateurs correspondant aux objectifs liés à la préservation des écosystèmes terrestres ou à la sécurité alimentaire par l'agriculture durable [(2)].

Pour ce qui est de la France, des expérimentations sont en cours dans trois domaines : l'exploitation des données issues des plateformes de location entre particuliers pour compléter les statistiques sur le tourisme (Franceschi, 2017), celle des enregistrements de téléphonie mobile pour la production de données sur la structure sociale des territoires, au-delà de ce que permettent déjà les sources administratives et le recensement, et enfin l'utilisation des offres d'emploi en ligne pour compléter la description du marché du travail. Ces statistiques expérimentales pourront apporter des informations inédites sur ces domaines émergents ou à mieux couvrir, mais sans pouvoir être immédiatement mises sur le même plan que des productions régulières éprouvées sur la longue période. L'apport de ces nouvelles données doit être testé au cas par cas : en extraire une information stable et conceptuellement cohérente n'a rien d'acquis, en particulier lorsqu'elles sont de type non structuré. Les Big Data ont un potentiel incontestable, la statistique publique cherche à en tirer parti, mais loin du mythe d'une réponse universelle et bon marché à la demande de statistiques toujours plus rapides, plus fiables et plus nombreuses.

(2) Voir le rapport d'un groupe de travail de l'ONU piloté par l'Australian Bureau of Statistics, https://unstats.un.org/bigdata/taskteams/satellite/UNGWG_Satellite_Task_Team_Report_WhiteCover.pdf

Références

BLANCHET D. & GIVORD P. (2017), « Données massives, statistique publique et mesure de l'économie », *L'Économie française*, édition 2017, collection Insee Références, pp. 59-77.

BORTOLI C. & COMBES S. (2015), « Apports de Google Trends pour prévoir la conjoncture : des pistes limitées », *Note de conjoncture*, mars, Insee, pp. 43-56.

BRYNJOLFSSON E., EGGERS F . & GANNAMAMENI A. (2017), « Using Massive Online Choice Experiments to Measure Changes in Well-being », draft, MIT.

CAVALLO A. & RIGOBON R. (2016), « The billion prices project: using online prices for measurement and research », *Journal of Economic Perspectives*, vol. 30, n° 2, pp. 151-178.

COHEN P., HAHN R., HALL J., LEVITT S. & METCALFE R. (2016), « Using Big Data to Estimate Consumer Surplus: The Case of Uber», *NBER Working paper* n° 22627.

FARRELL D. & GREIG F. (2016), « Paychecks, paydays and the online platform economy », JP-Morgan Chase Institute.

FRANCESCHI P. (2017), « Les logements touristiques de particuliers proposés par Internet », *Insee Analyse*, n° 33.

LAZER D., KENNEDY R., KING G. & VESPIGNANI A. (2014), « The parable of Google Flu : traps in big data analysis », *Science*, vol. 343 (6176), pp 1203-1205.

NATHAN M. & ROSSO A. (2013), *Measuring the UK's digital economy with big data*, rapport Growth Intelligence/NIESR.

OOSTROM L. *et al.* (2016), « Measuring the internet economy in the Netherlands : a big data analysis », *CBS working paper,* n° 2016-14.

SOES (2009), *CORINE Land Cover France Guide d'utilisation*, Document technique du Service de l'observation et des statistiques, Commissariat général au développement durable, ministère de l'Environnement.

Entretien avec Yves GASSOT

Propos recueillis par Edmond BARANES.

Yves Gassot a été pendant plus de vingt ans directeur général de l'IDATE DigiWorld. À ce titre, il a participé à de très nombreux travaux sur l'évolution des marchés des télécommunications, et plus largement du numérique, a assuré la direction de la revue *Communications & Stratégies*, du rapport annuel *DigiWorld Yearbook*, et il est l'auteur de nombreux articles. Il a été conseiller spécial de Viviane Reding, Commissaire européenne, pendant la revue du cadre réglementaire des communications électroniques. Il travaille aujourd'hui à la direction générale d'Orange. Yves Gassot est par ailleurs membre associé du Conseil général de l'Économie et du comité stratégique d'Iris Capital. Il a une formation initiale d'architecte (DPLG-Paris), une maîtrise d'urbanisme, et il est diplômé de l'IEP Paris (3e cycle d'aménagement).

***Enjeux numériques* [EN] : Qu'appelle-t-on Big Data ?**

Yves Gassot [Y.G.] : L'expression Big Data est née à la fin des années 1990 et est donc relativement récente. Elle est l'expression de plusieurs évolutions.

D'abord l'explosion de la production de données, qui combine la croissance continue du nombre d'internautes, amplifiée depuis dix ans par la généralisation des usages de l'Internet mobile, le basculement dans le numérique du monde de la musique et de la vidéo, l'intensification des échanges à travers les réseaux sociaux… Naturellement, l'entreprise est aussi partie prenante de la production de ces données, dans son fonctionnement interne comme dans ses relations avec les fournisseurs ou la clientèle. Enfin, il faut mettre en avant le phénomène émergent de l'Internet des objets (IoT) qui repousse à l'infini les limites des connexions et des flux supportés par l'Internet. Aussi, pour évaluer la production annuelle de données, on ne peut plus parler en gigaoctets (8 milliards de bits), ni même en téraoctets (1 000 gigaoctets) mais en dizaines de zettaoctets (1 000 milliards de gigaoctets [(1)]).

Derrière cette accélération de la production de données, on trouve évidemment des progrès multiples et considérables en matière de capacité de transmission (fibre optique, 4G, 5G et WiFi), de stockage, de calcul et traitement (à travers des composants plus performants, des infrastructures de *data centers* supportant des architectures en *cloud* et l'avènement de nouveaux algorithmes). La société britannique ARM, leader dans le design des composants pour smartphones, considère qu'entre 2009 et 2015, la performance des ordinateurs a été multipliée par 80, celle des cartes graphiques par 158 et celle de la résolution des écrans par 24…

[EN] : Peut-on parler de phénomène *data centric* ?

[Y.G.] : On parle aussi de « l'économie de la *data* ». Celle-ci se substituerait à l'économie de l'énergie associée à la révolution industrielle de la même façon que l'on voyait, dans les années 1990, la « société de l'information » succéder à la « société de consommation ». De fait, l'économie de la *data* domine aujourd'hui le *top ten* des plus grosses capitalisations boursières au niveau mondial (Apple, Amazon, Alphabet/Google, Microsoft, Facebook, mais aussi désormais Tencent, Alibaba…), en ayant bousculé les *majors* de l'énergie.

(1) On considère ainsi que 90 % des données disponibles aujourd'hui dans le monde ont été produites durant les deux dernières années.

Si, depuis plusieurs décennies, les systèmes d'information (SI) sont considérés comme le système nerveux des entreprises et des organisations, c'est plus récemment que le traitement de données en masse est apparu comme un facteur essentiel qui touche tous les aspects de transformation d'une entreprise. Ainsi, un opérateur de télécommunications cherchera dans les données un signe d'un possible désabonnement d'un client, mais aussi des automatismes pour optimiser en temps réel son réseau ou encore pour déceler une cyberattaque potentielle. Pour certaines entreprises, la donnée est intimement liée au modèle d'affaires. C'est naturellement le cas des grandes plateformes de l'Internet avec Google qui tire encore l'essentiel de ses recettes de la puissance et de la domination de son moteur de recherche tandis que Facebook arrive à un niveau de profitabilité record en commercialisant ses données sur le nouveau marché de la publicité ciblée (avec les risques associés, comme on l'a vu récemment avec les *fake news* et *data breach* de Cambridge Analytica).

Mais on ne doit pas oublier que la *data* trouve aussi sa place hors de l'économie purement marchande, avec le succès des sites d'échange (pas obligatoirement marchands), la gestion de la complexité urbaine (*cf.* les pratiques d' « open data » pour exploiter les multiples données urbaines afin d'optimiser les transports ou les plans d'urbanisme), et dans les progrès de la météorologie, de la médecine et plus généralement des activités scientifiques.

Dans le même temps, je crois qu'il est honnête de dire que le Big Data est surtout une pratique en devenir. Beaucoup d'entreprises, même importantes, sont encore dans ce domaine en phase d'apprentissage et ne le maîtrisent pas suffisamment pour l'intégrer dans le cœur de leurs activités. La migration vers le *cloud* est loin d'être achevée, l'IoT démarre… L'abondance de *data* peut s'assimiler à un certain chaos si l'on ne dispose pas des compétences (souvent « aspirées » par les leaders de l'Internet) et d'une stratégie sur les priorités de l'analyse et les données pertinentes à privilégier. Il faut dans ces conditions pouvoir compter sur une direction générale impliquée et prête à soutenir les objectifs mais aussi les aléas de la transformation numérique.

[EN] : Comment caractériser la rupture introduite par le Big Data *?*

[Y.G.] : Trois caractéristiques ont été associées au Big Data en mettant en avant la règle des « 3 V » :

- D'abord, et cela apparaît évident, il faut du Volume ;
- Il faut aussi de la Variété ; schématiquement, ce n'est pas en restant avec des données propres à une entreprise ou à un métier que l'on dispose des systèmes d'analyse les plus puissants et intéressants. D'où les nombreuses sociétés qui sont apparues en mettant en œuvre des compétences de traitement de la donnée mais aussi en intervenant comme *brokers* de *data* ;
- Enfin, c'est le plus souvent en temps réel que le traitement de ces données sera le plus pertinent (*Velocity*).

Mais si l'on parle de rupture, il faut aborder une autre notion, très directement associée au Big Data et qui même de plus en plus s'y substitue dans les commentaires sur la transformation numérique : l'intelligence artificielle ou ses derniers développements sous la forme de *machine learning*.

Ce sont en effet aujourd'hui les progrès fantastiques de l'intelligence artificielle qui peuvent donner de nouvelles ambitions et perspectives au traitement et à la valorisation de la donnée.

L'intelligence artificielle est une expression assez ancienne qui est apparue dès les années 1950 dans les laboratoires de Stanford. On l'a ensuite un peu oubliée, même si l'on a connu, dans les années 1970, l'émergence de systèmes experts. Il s'agissait alors de décortiquer des sujets, pas trop complexes, en leur appliquant des règles de causalité établies sur la base d'un jeu limité de données structurées. Avec le Big Data, on est passé très schématiquement de cette approche causale à une recherche des corrélations à travers l'utilisation d'un très grand nombre de données (souvent non structurées). Les développements sont en particulier focalisés sur les algorithmes basés sur une

méthode d'apprentissage (*machine learning*) qui va permettre à une machine de maîtriser progressivement les multiples combinatoires d'une partie de Go, ou de distinguer l'image d'un chat de celle d'un chien… Mais malgré les performances spectaculaires obtenues et leur importance pour les domaines de perception (reconnaissance des formes et reconnaissance vocale), les experts du domaine insistent sur le fait qu'il s'agit toujours d'une intelligence supervisée. Il a fallu que des milliers d'images soient codées comme étant celles de chiens pour que la machine apprenne à faire la différence avec celles des chats, tandis que la machine de *deep mind* (Alphabet/Google), qui a battu en 2016 le champion du monde de Go, ne saurait pas s'aventurer sans « apprentissage supervisé » dans une partie d'échecs…

On semble donc encore loin d'une intelligence artificielle au sens où elle disposerait de l'intelligence humaine de perception et de raisonnement, sans même parler de conscience.

[EN] : Quels sont les nouveaux enjeux réglementaires ?

[Y.G.] : À côté des progrès attendus, il y a en effet un ensemble de problèmes réglementaires très complexes. On se contentera d'en citer quelques-uns :

- un risque sur la préservation des données personnelles : la collecte des données qui contribue au financement de l'innovation et des applications offertes aux internautes s'accompagne de risques de cybercriminalité et de craintes sur la maîtrise des données personnelles (*cf.* le règlement européen dit de RGPD qui entre en application en mai 2018) ;
- un problème de concurrence : la domination des grandes plateformes de l'Internet est alimentée par l'accumulation continue de données de par leur modèle et par la puissance financière acquise (qui leur permet d'acheter les talents et les *start-ups* les plus prometteuses dans le domaine de l'intelligence artificielle).

Ainsi, comme souvent, la concentration de la technologie au sein d'un petit nombre d'acteurs s'apparente à un phénomène complexe à réguler car il recouvre :

- des avantages pour le consommateur ainsi que pour les autres acteurs de l'industrie qui trouvent l'occasion de s'inscrire dans des écosystèmes performants ;
- mais aussi des risques sur le contrôle de nos données personnelles et la distorsion du marché par de nouvelles barrières à l'entrée.

Si l'on prend les interfaces vocales, basées sur les derniers progrès de l'intelligence artificielle, qui sont installées dans les smartphones ou les enceintes domestiques par Apple, Google, Amazon, Samsung et Microsoft, le consommateur peut considérer un progrès sans percevoir tout de suite que ce mode de commande peut aussi s'accompagner d'une limitation du choix offert pour accéder à tel ou tel catalogue de musique en *streaming*. On peut ainsi craindre que ces dernières innovations associées à l'intelligence artificielle viennent avantager les applications des principales plateformes mais aussi freiner le changement de plateformes par le consommateur…

De plus, les dispositions réglementaires esquissées au niveau d'un pays ou d'un groupe de pays (Union européenne) doivent, pour être efficaces, être négociées dans un cadre international, alors même que les enjeux se présentent souvent différemment pour les différents acteurs. On peut à cet égard évoquer les négociations à répétition entre l'Europe et les États-Unis à travers le Safe Harbor puis le Privacy Shield, mais aussi les discussions entre les membres de l'Union européenne qui n'ont pas tous la même sensibilité sur ce sujet.

Il me paraît aussi important de mentionner d'autres débats, par exemple :

- celui qui s'intéresse à la symétrie des règles à l'intérieur de la chaîne de valeur ; ainsi les opérateurs de télécommunications soulignent que les fournisseurs de service Internet (OTT) seraient beau-

coup moins encadrés dans la gestion des données personnelles que les opérateurs de télécommunications par leur régulation sectorielle ;

- celui qui oppose de plus en plus souvent les grands acteurs de l'Internet et les autorités publiques ; on a vu Microsoft s'opposer au droit de regard des autorités américaines sur des données inscrites dans un de ses *data centers* en Irlande ; Apple a refusé, dans une affaire criminelle aux États-Unis, d'« ouvrir » le système d'exploitation d'un iPhone ; l'application Whatsapp a été (momentanément) fermée au Brésil sur la requête d'un juge qui souhaitait avoir accès au contenu de cette messagerie cryptée.

D'un autre côté, il faut aussi souligner le poids que peut avoir la régulation « par la *data* », en faisant notamment appel aux consommateurs dans l'esprit des réflexions de Nicolas Colin et Henri Verdier (*L'Âge de la multitude*). Le régulateur français des télécoms (ARCEP) s'est engagé dans cette voie en mettant des cartes géographiques à disposition des usagers pour leur permettre de noter la qualité des services mobiles sur le territoire.

[EN] : Faut-il réguler les plateformes et s'assurer de la loyauté des algorithmes ?

[Y.G.] : Je pense qu'il est assez difficile et peut-être dangereux de prescrire une réglementation *ex ante* à vocation générale mais propre aux plateformes. Il faut peut-être d'abord s'assurer de ce que les différents droits (droits des consommateurs, droits des contrats, droits intellectuels, droit fiscal, droit de la concurrence…) soient respectés, et quand c'est nécessaire construire les rapports de force qui permettent d'imposer les remèdes pertinents. Des progrès sont possibles et cela a commencé. On a vu que la puissance publique pouvait imposer à une grande plateforme de réservation l'abandon du principe d'exclusivité qu'elle imposait aux hôtels. On a également vu que Google a dû amender le fonctionnement de son moteur de recherche. On peut imaginer contrôler la loyauté des algorithmes utilisés par les plateformes dominantes… par exemple, en s'assurant de temps en temps par sondage que les recommandations d'un site d'eCommerce vous informant de ce que vos amis ont aimé …. sont réellement fondées !

Cela étant, je suis très inquiet quand je vois que le premier réflexe devant une innovation de rupture se limite à la réglementation. Il faut bien sûr aussi compter sur la dynamique d'innovation et de concurrence, et pour cela favoriser les investissements dans la connaissance et la création.

Après tout, on a vu apparaître de nouvelles plateformes sectorielles (Uber, AirBnB, Spotify, Netflix…) et les GAFAM [(2)] peinent souvent à réussir leur diversification « horizontale »…

[EN] : Quel constat peut-on faire sur le positionnement de l'Europe face au dynamisme des États-Unis et au rattrapage de la Chine ?

[Y.G.] : L'Europe est absente des dix plus grosses capitalisations boursières qui sont, comme je l'évoquais, dominées par les GAFAM et désormais leurs homologues chinois (Tencent, Alibaba, Baidu). La situation n'est donc pas brillante mais il faut bien voir que nous ne sommes qu'au début de la transformation numérique. Nous conservons des atouts, dans l'IoT, dans l'intelligence artificielle avec des chercheurs reconnus dans le monde entier, dans les télécommunications, dans les industries culturelles (notamment les jeux vidéo). Bien que les investissements européens en capital-risque soient encore en retrait, l'écart avec les États-Unis se réduit et les écosystèmes de *start-ups* à Londres, Paris ou Berlin vont certainement générer de plus en plus de licornes… Enfin, la transformation numérique, en touchant des secteurs de l'industrie dans lesquels l'Europe dispose d'acteurs puissants (aéronautique, automobile, énergie…), offre de nouveaux challenges à notre industrie (Industry 4.0) et des terrains d'émergence de nouveaux leaderships européens dans le numérique.

(2) Google, Apple, Facebook, Amazon et Microsoft.

[EN] : Quelles seraient vos recommandations de politiques à mettre en œuvre au niveau européen ?

[Y.G.] : 1. L'éducation ‑ par la qualité et la profondeur de la formation initiale, supérieure et « tout au long de la vie » ‑ constitue une priorité connue mais qui n'est pas encore assez prise en compte. Elle doit s'accompagner d'un renforcement de la culture scientifique sans qu'on fasse une mauvaise querelle aux sciences « molles ». C'est la condition pour ne pas subir la transformation numérique, mais aussi pour éviter que l'Europe subisse la polarisation des emplois.

2. Il faut poursuivre les progrès observés dans le domaine des *start-ups* et de la culture d'entrepreneur, adopter les dispositions fiscales permettant de mieux orienter l'épargne pour renouveler notre tissu économique et monter en gamme. C'est en réduisant le *gap* de productivité entre les 10 % des entreprises les plus avancées dans l'intégration numérique et celles qui restent encore très en retrait que l'on évitera la fameuse « stagnation séculaire ».

3. Malgré les vicissitudes du projet européen, les pays membres doivent poursuivre leurs efforts pour disposer d'un cadre réglementaire harmonisé afin de bénéficier d'un grand marché domestique et de favoriser l'émergence de leaders européens. Mais l'Europe du numérique doit aussi s'exprimer comme un projet mobilisateur, alternatif aux dérives qui peuvent ruiner la confiance du plus grand nombre. C'est l'enjeu des discussions sur le régime fiscal des plateformes, sur les abus de position dominante, ainsi que sur la maîtrise des données personnelles et la lutte contre la cybercriminalité.

4. La valorisation de la place éminente du numérique dans les économies d'énergie (pas de systématisation des énergies alternatives sans « réseaux intelligents ») et la diminution des pollutions (*smart cities*) doit enfin être renforcée et devenir un élément de confiance dans l'innovation et l'adhésion à un axe stratégique associant transformation numérique et transition énergétique.

[EN] : Finalement, quelle est votre vision sur l'évolution du Big Data dans les prochaines années ?

[Y.G.] : Quand on observe les différentes catégories dans lesquelles se rangent les *start-up*s ainsi que les technologies mises en œuvre, on s'aperçoit de ce que le Big Data, combiné avec les progrès de l'intelligence artificielle, est une source d'innovation qui s'applique à tous les secteurs de l'innovation numérique. L'IoT va permettre de drainer de multiples données et donc de transformer radicalement la maintenance des grands systèmes complexes (« maintenance prédictive »). La cybersécurité, en intégrant de plus en plus d'intelligence artificielle, va pouvoir détecter des « signaux faibles » annonciateurs d'attaques. Le traitement du cancer ne pourra plus se concevoir sans le concours de systèmes capables d'exploiter de gigantesques bases de données pour préciser le diagnostic et évaluer le meilleur traitement. Et l'on pourrait continuer en faisant référence à la voiture autonome, à la virtualisation des réseaux de télécommunications, à la FinTech et l'InsurTech, à la publicité programmatique, etc.

Les perspectives sont donc très impressionnantes mais encore une fois, nous n'en sommes qu'aux balbutiements… et les usages concrets dans les entreprises sont encore parfois très sommaires. Le *cloud*, l'IoT sont encore en phase de déploiement. Les options réglementaires sont encore incertaines. La place que va prendre la cybercriminalité n'est pas bien cernée… Dans le contexte actuel, le rythme auquel se généralisera le Big Data, et surtout les scénarios « industriels » et géopolitiques qui l'accompagneront, ne sont pas encore très clairs.

Hors dossier

Compte-rendu de la Journée 2017 du Conseil scientifique de l'AFNIC (Association française pour le Nommage Internet en Coopération)

Depuis sept ans, l'Association française pour le Nommage Internet en Coopération (AFNIC), gestionnaire historique du .fr, organise les Journées du Conseil scientifique de l'AFNIC (JCSA). Cet événement permet de faire un point sur les avancées en termes de recherche et standardisation sur les protocoles utilisés dans l'Internet, et plus particulièrement le *Domain Name System* ou DNS. Le DNS permet principalement de faire le lien entre les noms des équipements et les adresses utilisées par ceux-ci sur le réseau. Cette brique est fondamentale pour le fonctionnement de l'Internet, elle masque un objet technique comme une adresse IP (192.134.5.24) au profit d'un nom plus facilement exploitable par un être humain (afnic.fr). Au cours du temps, les fonctions du DNS ont été étendues, par exemple pour permettre de répartir la charge des sites web les plus fréquentés, de gérer des clés de chiffrement, d'enregistrer d'autres identifiants que les noms d'équipements... L'architecture du DNS est en constante évolution, pour à la fois augmenter les performances face à la croissance du nombre de requêtes, renforcer la sécurité de ce service critique et donc réduire l'impact des attaques, mieux protéger la vie privée, et ajouter des fonctionnalités en exploitant l'infrastructure robuste déjà déployée. Les membres du conseil scientifique étant proches des milieux universitaires et de la standardisation, ils contribuent à la réflexion de l'AFNIC sur les moyens à mettre au service de ses missions. Ils se prononcent sur les grandes orientations en matière de recherche et développement, de veille technologique et de gouvernance de l'Internet.

Après s'être intéressé les années précédentes à l'Internet des objets, aux architectures alternatives, à la métrologie et à la résilience, le conseil scientifique de l'AFNIC a traité le 6 juillet dernier de la gestion de la vie privée. Depuis la création du DNS dans les années 1980, les requêtes DNS successives menant à la résolution d'un nom n'ont jamais été chiffrées, et de plus elles remontent jusqu'à la racine. Des serveurs relais mis en place pour limiter le nombre d'interrogations masquent généralement l'identité de l'utilisateur, mais il peut arriver que l'information sorte d'un domaine ; pour une entreprise, cela peut conduire à une fuite d'informations confidentielles, comme par exemple la liste de ses fournisseurs. Ces problèmes peuvent être exacerbés, si les noms à résoudre contiennent des informations sensibles (par exemple, si ceux-ci incluent des identifiants obtenus en scannant des qr-codes, leur interception peut conduire à dévoiler un procédé de fabrication).

La JCSA a fait le point sur les différentes approches mises en œuvre par l'IETF (Internet Engineering Task Force), l'organisme international qui standardise les protocoles de l'Internet, pour diminuer les vulnérabilités du protocole liées à la gestion de la vie privée. Deux améliorations ont été proposées.

La première se base sur la nature hiérarchique des noms de domaines. Quand un utilisateur cherche à résoudre le nom « qrcode.ident.exemple.fr », la requête est envoyée à un résolveur qui est situé dans l'entreprise, chez son fournisseur d'accès ou chez un résolveur dédié (dit « ouvert ») comme en offrent Google, Cisco, etc. Celui-ci va interroger les serveurs « racines » du DNS pour

localiser les serveurs gérant la zone .fr, puis interroger ces serveurs pour localiser exemple.fr, et ainsi de suite, jusqu'à obtenir la réponse à la question.

Dans l'approche initiale, l'intégralité de la requête est fournie aux différents serveurs. Avec les extensions pour la gestion de la vie privée (dites « minimisation des requêtes »), seule la partie intéressant le serveur lui sera envoyée. Ainsi le serveur racine ne recevra qu'une requête concernant .fr, au lieu de qrcode.ident.exemple.fr

La deuxième amélioration concerne le chiffrement des requêtes entre la machine et le résolveur. Le changement majeur est le passage du protocole UDP (sans état) au protocole TCP (avec état) pour pouvoir utiliser le protocole de chiffrement TLS, utilisé également pour sécuriser les échanges du web. C'est de cette deuxième amélioration impliquant des changements protocolaires et comportementaux qu'a traité la JCSA.

La matinée de la JCSA est généralement dédiée à des tutoriaux. Sara Dickinson de Sinodun et Willem Toorop de NLnet Labs ont présenté et fait une démonstration du logiciel getDNS et de ses interfaces de programmation incluant les extensions pour la gestion de la vie privée. Maxence Tury de l'Agence nationale de la Sécurité des Systèmes d'Information (ANSSI) a présenté le fonctionnement du protocole TLS et, par le biais d'exemples interactifs avec le logiciel Scapy dans des machines virtuelles, a expliqué comment les certificats peuvent être utilisés pour chiffrer les communications.

L'après-midi a été consacré à des exposés des différentes technologies. Sarah Dickinson est revenue plus en détail sur les travaux de l'IETF dédiés à la gestion de la vie privée pour le DNS, et Alexander Mayrhofer, du registre nic.at, a présenté les premières études sur le dimensionnement des serveurs utilisant TLS. Il s'avère que les changements proposés auront peu d'impact sur les performances du DNS. Marck To, de la société EfficientIP, a ensuite réalisé une démonstration montrant comment des données peuvent être exfiltrées d'un site en utilisant le DNS, et quelles sont les méthodes pour se protéger de cette attaque. La dernière présentation du séminaire portait sur la thématique plus générale de la future application RGPD (Règlement général sur la Protection des Données de l'Union européenne), et son impact sur la conception des services a été abordé par Bruno Rasle de AFCDP.

L'ensemble des conférences de la journée (vidéos et supports de présentation) est accessible en ligne sur
https://www.afnic.fr/fr/l-afnic-en-bref/actualites/actualites-generales/10660/show/jcsa17-retour-sur-l-edition-2017-de-la-journee-du-conseil-scientifique-de-l-afnic.html

Hors dossier

La prochaine révolution est celle des émotions

Par Laure KALTENBACH
Co-fondatrice de CreativeTech

Intelligences artificielles, robots, *cobots*, drones, algorithmes, réalités virtuelle, mixte ou augmentée, hologrammes, objets connectés : nous basculons dans un monde où humanité et technologie se combinent, s'augmentent, se confrontent ou s'opposent. On ne sait plus très bien.

Alors que ces mots envahissent les écrans et les esprits, les enjeux sont loin d'être décodés : un sondage cité par *Sciences et Avenir* en septembre 2017 indique que « 34 % des Français disent ne pas savoir ce qu'est l'intelligence artificielle, 41 % croient le savoir et 25 % pensent le savoir précisément ».

Le point de bascule

« Le pouvoir de l'homme s'est accru dans tous les domaines, excepté sur lui-même », disait Churchill. Ces mutations technologiques semblent le contredire tant il est vrai qu'elles explorent sans relâche nos cinq sens, nos émotions, sollicitant l'imagination et les interactions dans des univers de 0 et de 1 : une vue décuplée et augmentée, une ouïe diésée et spatialisée, un toucher hypersensible – désormais avec une capacité de projection du corps –, un odorat reproductible. Seul le goût laisse encore – mais pour combien de temps ? – à désirer.

Là où les générations précédentes confinaient l'émotion au cercle de l'intime et la tenaient pour une sensiblerie, voire une faiblesse, nos contemporains annoncent au contraire l'avènement du règne des émotions, avec force réseaux sociaux et émoticônes [(1)].

Cette révolution technologique qui tente de conquérir les *terrae incognitae* de nos émotions coïncide avec deux autres transformations décisives.

La réactualisation des sciences cognitives et en particulier des neurosciences révolutionne la compréhension des émotions. « L'émotion est d'abord une révolution scientifique : nous disposons désormais des outils pour montrer que la conscience est holistique et qu'elle est sensoriellement sans limite », indique la philosophe et psychanalyste Cynthia Fleury. La dernière transformation majeure est politique, qui voit s'affronter le rationnel et l'émotionnel dans la démocratie. Cynthia Fleury précise : « Nous assistons à une révolution démocratique sur la place des émotions dans les institutions, dont le rôle est précisément d'instiller de la rationalité alors que les individus sont submergés d'émotions. »

Ces trois accélérations, technologique, scientifique et politique, interrogent les acteurs publics dans leur appréhension des questions éthiques et réglementaires – le droit des robots est d'ores et déjà en débat. Elles questionnent également frontalement les entreprises dans leur compréhension de leurs clients et collaborateurs, ainsi que dans la conception même de leurs produits et services. La révolution des émotions est donc en marche.

(1) Aussi appelés *smileys*.

Nous sommes entrés dans l'ère de l'économie de l'émotion

« Bien contrôlée, l'intelligence artificielle pourrait nous conduire à devenir encore plus humains, un hyper-humanisme », annonce Joël de Rosnay. Encore plus humains ? Encore plus sensibles aux équilibres naturels qui nous entourent ? Les acteurs publics sont ainsi confrontés à « un nouvel ordre du jour humain » pour reprendre les propos de l'historien Yuval Noah Harari. Toutes les problématiques sont convoquées : éthique, réglementation, recherche et développement, gouvernance.

Comment penser l'éthique avec les *data* et « machines conscientes » ? Différentes initiatives voient le jour depuis quelques années. En 2014, une déclaration préliminaire des droits de l'humain numérique a été élaborée (www.ddhn.org), pour que la *soft law* préempte ce futur aléatoire. Pour Laurence Devillers, professeure à Paris-Sorbonne et spécialiste de l'*affective computing*, plusieurs approches coexistent pour aborder les questions éthiques de l'intelligence artificielle (IA) : l'éthique d'abdication – l'IA est programmée pour s'autodétruire en cas de problème, et par exemple, dans le cas d'un véhicule autonome, risquer la vie de ses occupants –, l'éthique déontique – l'IA applique *stricto sensu* les règlements –, l'éthique conséquentialiste – l'IA arbitre en fonction de statistiques, par exemple le nombre de morts dans le cas d'un véhicule autonome. Quelle combinaison d'éthiques devons-nous dessiner pour cohabiter avec les intelligences artificielles ? Sur un ton plus léger, le « suicide », le 27 juillet 2017, dans une fontaine d'un centre commercial de Washington DC, d'un robot de sécurité, pourtant conçu pour ne ressentir aucune émotion, interroge.

Il pose également la question du droit des robots. La question peut sembler à première vue ubuesque. Créée en 2014 par l'avocat Alain Bensoussan, l'Association du droit des robots appelle à la création d'un cadre juridique propre à la robotique, à l'instar du droit de l'informatique ou des télécommunications. Pour ses promoteurs, le droit des robots est inéluctable, à mesure que les machines se dotent d'intelligence artificielle les autonomisant chaque jour un peu plus.

Le soutien à la recherche est naturellement au cœur du sujet. Avec plus de 5 000 chercheurs en intelligence artificielle, la France dispose d'une force de développement réelle. Si la recherche est bien dotée, les usages sont quant à eux moins étudiés. Or, comme pour les précédentes révolutions technologiques, ils sont un enjeu crucial pour nos économies. Pour détecter et inventer de nouveaux usages, Jean-Gabriel Ganascia, professeur en informatique à l'Université Pierre-et-Marie-Curie et chercheur en IA, plaide pour un renforcement de la pluridisciplinarité, avec des sociologues notamment.

Les accélérations fulgurantes décrites par Yuval Harari dans sa *Brève histoire de l'avenir* interpellent : « Au XXIe siècle, les principaux produits de l'économie ne seront plus les biens matériels mais les corps, le cerveau et la conscience, autrement dit la vie artificielle. L'*Homo deus* – l'homme devenu dieu – a trois façons de passer au niveau supérieur : la bio-ingénierie, les cyborgs et la vie anorganique. » Comment penser un cadre pertinent pour penser ces enjeux ? Ces derniers mois, les initiatives se multiplient outre-Atlantique : fonds de financement dédiés (création du Salesforce AI Fund, fonds Google...), création du *think tank* The Future Society de la Harvard Business School, qui souhaite ouvrir à tous le débat autour de l'IA (ai-initiative.org), prises de parole spectaculaires des grands acteurs numériques (Elon Musk, Mark Zuckerberg, Bill Gates, Larry Page, Jeff Bezos...). En France, les lignes bougent également : mission d'information parlementaire sur l'IA confiée à Cédric Villani ; proposition du président de la République, dans son discours du 26 septembre à la Sorbonne, de créer une Agence européenne de l'Innovation pour financer, notamment, l'intelligence artificielle. Mais ne faut-il pas urgemment marquer les esprits et mettre en place une *Conference of Parties* sur l'intelligence artificielle ? Faire preuve d'audace avec une « CoPIA », qui réunisse tous les acteurs impliqués dans cette nouvelle économie de l'émotion : chercheurs, entrepreneurs, artistes, ingénieurs, philosophes, sociologues. Une CoPIA qui renou-

velle le genre, ouverte à toutes les disciplines, une CoPIA post-innovation ouverte : les solutions de demain ne peuvent être laissées aux seuls spécialistes techniques. En effet, Laurence Devillers nous prévient : « Il faut travailler les discriminations qui pourraient intervenir dans le domaine de l'intelligence artificielle. Va-t-on reproduire en IA les clivages à l'œuvre dans la société ? On s'aperçoit que les femmes-agents conversationnels sont principalement utilisées comme assistantes de soins, qu'il risque d'y avoir à profusion des robots sexuels de genre féminin (…) ou que les logiciels de reconnaissance de visage, comme celui de Google, savent mieux identifier les peaux blanches que les noires, faute d'avoir assez de données dans leur corpus d'origine. Il faut avoir la réflexion de fond sur ces sujets pour ne pas arriver à la bêtise artificielle. »

Les auteurs d'émotion

Pour les entreprises, une nouvelle bataille s'engage : introduire de l'émotion dans les processus d'innovation. Puisque nos émotions dictent de plus en plus nos comportements et nos actions, les entreprises doivent elles aussi parler le langage des émotions. Management émotionnel, intelligence émotionnelle, marketing émotionnel, relation émotionnelle… Aucun terrain ne semble oublié, et le pouvoir créatif des artistes est déjà largement mis à contribution pour créer une publicité, designer un objet, ré-enchanter un espace ou un lieu…

Mais à y regarder de près, il y a, à ce jour, un pan entier de l'activité d'une entreprise qui intègre encore très peu la dimension émotionnelle et la distinction créative : c'est la conception et la réalisation des produits et services eux-mêmes. Or, un constat s'impose : les produits se ressemblent de plus en plus tant pour la fiabilité que pour les fonctionnalités ou les prix. Difficile dans ces conditions de se distinguer. La différence peut se faire sur l'emballage, la communication, ou plus profondément sur les valeurs de l'entreprise, l'identité de la marque… Mais surtout, la différence va se faire de plus en plus sur la prise en compte d'une dimension émotionnelle au cours du processus de développement de nouveaux produits et services, en l'intégrant à la méthodologie de travail des ingénieurs, chercheurs et techniciens. C'est donc très en amont, y compris dans la gestion interne des collaborateurs, que la bataille commerciale va se jouer entre les entreprises. L'émotion est ce qui donne au produit ou au service sa singularité et son irréductible sensibilité. L'appel à la pluridisciplinarité de Jean-Gabriel Ganascia doit s'appuyer sur différents talents. L'intégration des talents artistiques dans les processus d'innovation reste à inventer.

Les entreprises ne favorisent pas spontanément la créativité, qui remet en cause les habitudes, les acquis, les traditions, perturbe les organisations. Produire du nouveau en permanence est source de désordre. Et pourtant, dans une société de l'émotion, la créativité devient une qualité primordiale. Les artistes construisent et créent à travers les questions posées à nos sociétés : qu'ils soient scénaristes, scénographes, vidéastes, plasticiens, musiciens, danseurs, cinéastes, concepteurs de jeux vidéo… Le lien particulier des artistes avec la société, leur approche du sensible, leur esthétique, leur vision forment une chance qui va bien au-delà de la pluridisciplinarité des équipes pour concevoir de nouvelles idées. Les découvertes d'Antonio Damasio, professeur de neurologie, neurosciences et psychologie, directeur de l'Institut pour l'étude neurologique de l'émotion et de la créativité de l'Université de la Californie du Sud, sont à cet égard exemplaires. Spécialiste des liens entre la conscience, les émotions, les sentiments et le corps, féru d'art, Antonio Damasio rapproche neurosciences et art : « Il y a deux façons de comprendre l'univers : l'approche scientifique et la démarche artistique. Elles n'ont ni les mêmes caractéristiques, ni la même approche, mais elles permettent toutes deux de mieux connaître la condition humaine. Mais l'art, plus que la science, et c'est leur différence, est lié aux émotions (…). Pour un artiste, c'est donc très important de connecter l'imagination aux émotions. »

Antonio Damasio et les neurosciences nous laissent entrevoir, sous un nouveau jour, toute la place de l'émotion dans le processus créatif. Encore faut-il s'emparer de ces découvertes. Tout simplement parce que la capacité des artistes à rendre uniques, singulières et… pourtant universelles leurs œuvres résonne avec la volonté des entreprises de trouver des produits et services uniques, singuliers … et pourtant universels.

L'entreprise est déjà sortie de sa tour d'ivoire en se lançant dans l'*open innovation*, une manière pour elle de mobiliser la capacité d'innovation « hors les murs ». Mais elle n'a, ce faisant, parcouru que la moitié du chemin. Il lui faut maintenant aller un cran plus loin, en se lançant dans l'*open creativity*. C'est à cette condition que l'entreprise sera capable de recevoir ce que les artistes peuvent apporter : le détournement d'une idée, d'une innovation, d'une technologie pour lui donner un « supplément d'âme », lui faire porter du sens. C'est précisément ce qui va permettre à l'entreprise de devenir une *creative tech*, c'est-à-dire une entreprise connectée au nouveau monde, un monde qui conjugue technologies et émotions.

Le mot « robot » lui-même, issu des langues slaves - corvée, travail forcé -, serait apparu pour la première fois dans les années 1920 dans la pièce de théâtre *R.U.R (Rossum's Universal Robots)* de l'auteur tchèque Karel Capek qui l'aurait lui-même emprunté à son frère Josef, peintre et écrivain. Laissons donc la place aux talents artistiques, sous toutes leurs formes, pour qu'ils nous aident, aux côtés des chercheurs, ingénieurs, sociologues, philosophes, entrepreneurs et acteurs publics, à reprendre le chemin des Humanités, version XXI^{e} siècle.

Résumés

Les concepts mathématiques et algorithmiques à l'œuvre dans l'apprentissage et la capacité prédictive des machines ne sont pas très nouveaux, mais ils trouvent désormais une utilisation massive avec l'explosion de la quantité disponible de données. Le Big Data attire et fait peur en même temps, les risques qui lui sont associés ne pourront être maîtrisés que si la culture probabiliste et statistique particulière à ce domaine se diffuse bien au-delà du petit monde des *data scientists*.

Cet article vise à mieux comprendre comment s'organisent actuellement les échanges de données. Nous commençons par décrire le côté de la demande, en étudiant pourquoi, et comment, les données acquièrent de la valeur. Nous considérons ensuite le côté de l'offre, en nous demandant d'où viennent les données et qui en contrôle la production et la collecte. Enfin, nous décrivons les différentes modalités sous lesquelles l'offre et la demande se rencontrent. Nous constatons qu'une quantité sans cesse croissante de données est produite, collectée et utilisée mais qu'en définitive, une fraction assez limitée de ces données est échangée. Nous proposons trois explications : le caractère stratégique des données pour les entreprises, la difficulté d'organiser des places de marché décentralisées et le manque de contrôle des individus sur les données qu'ils produisent.

Les données personnelles prennent une place de plus en plus importante dans le positionnement stratégique des entreprises de l'Internet en leur permettant de toujours mieux cibler les consommateurs. Quand ces données sont combinées avec d'autres données, par exemple des données administratives, leur exploitation peut générer une valeur ajoutée unique pour les entreprises. Nous soutenons que ces nouvelles stratégies qui visent à extraire une valeur des données personnelles justifient une régulation appropriée du marché. Premièrement, il apparaît important d'identifier les sources de valeur liées à l'exploitation des données personnelles. Deuxièmement, nous mettons en évidence, en nous appuyant sur la littérature académique en économie et en marketing, les stratégies de valorisation des données personnelles auxquelles les entreprises peuvent avoir recours, et la manière dont de nouveaux modèles d'affaires peuvent être ainsi stimulés. Troisièmement, nous nous interrogeons sur le rôle de la régulation, qui vise à protéger la vie privée des individus tout en préservant la capacité d'innover des entreprises.

des technologies et de la société conduisent à préciser comment les armées conçoivent l'exercice de leur mission dans l'espace numérique. Avant d'expliciter le rôle du ministère en tant qu'acteur, il convient de décrire la compréhension de la souveraineté nationale sous le prisme de la défense. La première partie de cet article est ainsi consacrée à la construction d'une définition efficace de la souveraineté numérique pour les armées. La suite présente l'ambition numérique du ministère des Armées, l'enjeu d'acquisition des connaissances et de l'anticipation, et enfin les capacités d'actions conservées en propre pour répondre à ses attributions.

Abstracts

rities to private data, attention is turned to the grounds underlying this trend and the obstacles in its way. A grid is proposed for analyzing the provisions for a regulation that would protect the interests of the parties involved while allowing for data to circulate and be shared for the common interest.

Digital technology and big data are not the cure-all claimed by Silicon Valley. They also cause much harm, or even make the problems and defects of our societies worse. To regulate big data, a European ethics with respect for individuals and for democratic institutions must oppose the ethics of the Far West.

The collection, analysis and presentation of data as proof in a penal case are the very grounds for the fight against criminality. The concrete examples cited in matters of cybercriminality and intelligence demonstrate the need to exercise control over data.

The Ministry of the Armed Forces is a major figure in the exercise of national sovereignty in the digital realm. The French Constitution legitimates the armed forces as the ultimate defense of the state's sovereignty. Cyberspace is no exception. Rapid changes in technology and society are forcing us to clarify the armed forces' conception of their assignment in cyberspace. Before describing the Ministry's role, the concept of national sovereignty is discussed in terms of defense. The Ministry of the Armed Forces is seeking, in digital matters, to acquire knowledge for anticipating events and the capacity to undertake actions to fulfill its duties.

What impact do big data have on business-to-business (B2B) marketing? After a brief recall of recent trends in big data and data science, focus is shifted onto the data about firms and the situation in France. How is the big data revolution modifying two major marketing trends (marketing studies and the management of the "commercial funnel")?

The French national electricity company (EdF) has a fleet of operational nuclear power stations that counts 58 reactors built in standardized stages and, on the average, 32 years old. It also has a legacy of abundant, diverse data that could be processed using new techniques in data analytics that are part of the big data bang. Following a phase for demonstrating these techniques, ÉdF has generalized the tools of data analytics and procedures for improving the company's data legacy. This combination is a lever for improving the safety, operations and maintenance of this fleet of nuclear reactors.

MISCELLANY

The dialog between technology and creativity is at the dawn of a new revolution that is predicted to be intellectual, imaginative, romantic and emotional, and that will give a surprising texture and subtlety to sex, to borrow from Anaïs Nin's *Delta of Venus*. How long will we have to wait for an algorithm in a seductive attire to be uplifting for our five senses? Spike Jonze's movie *Her* partly responds to this; and *Blade Runner 2049* has given form to desirable forms of artificial intelligence. Many another is trying…

Ont contribué à ce numéro

François BANCILHON, diplômé de l'École des Mines de Paris, titulaire d'un PhD de l'Université du Michigan et d'une thèse d'État de l'Université de Paris-XI, a eu une double carrière : une première dans la recherche académique (chercheur à l'INRIA et MCC, professeur à l'Université de Paris XI), et une deuxième dans l'industrie. Il a cocréé ou dirigé plusieurs entreprises (O_2 Technology, Arioso, Xylème, Ucopia, Mandrakesoft/Mandriva et Data Publica/C-Radar). Il a partagé sa vie professionnelle entre la France et les États-Unis. Il est actuellement directeur des projets innovants chez Sidetrade, une société qui développe et commercialise une plateforme SaaS d'intelligence artificielle qui redessine l'engagement client.

Edmond BARANES est professeur d'économie à l'Université de Montpellier. Ses principaux thèmes de recherche sont l'économie industrielle, la politique de la concurrence et la réglementation en la matière. Ses travaux s'intéressent en particulier aux marchés de l'économie numérique, de l'énergie et de la santé.

Paul BELLEFLAMME est titulaire d'un doctorat en économie de l'Université de Namur (1997). Son parcours professionnel de professeur d'économie l'a mené de Queen Mary University of London à l'Université catholique de Louvain, en passant par Aix-Marseille Université et Kedge Business School. La recherche et l'enseignement de Paul Belleflamme s'inscrivent dans les champs de l'organisation industrielle et de l'économie de l'innovation, avec un intérêt particulier pour le monde du numérique (auquel il dédie son blog, www.IPdigIT.eu).

Véronique BELLON-MAUREL, ingénieur des Ponts, des Eaux et des Forêts, est directrice du département Écotechnologies à l'Irstea et directrice de l'institut Convergences Agriculture Numérique #DigitAg à Montpellier.

Didier BLANCHET est directeur des études et synthèses économiques à l'Insee. Sa direction est notamment en charge de la production des comptes nationaux et de la collecte des enquêtes de conjoncture. Né en 1957, il est diplômé de l'École polytechnique et de l'ENSAE, titulaire d'une thèse de doctorat et d'une habilitation à diriger les recherches de l'Institut d'Études politiques de Paris. Il a débuté sa carrière comme chercheur à l'Institut national d'Études démographiques. Il a rejoint l'Insee en 1993. Il y a été successivement chef de la division Redistribution et politiques sociales, directeur de l'ENSAE, chef du département Emploi et Revenus d'activité, chef du département des Études économiques d'Ensemble, rédacteur en chef de la revue *Économie et Statistique*.

Grazia CECERE est une professeur d'économie à Télécom École de Management à Paris, école qui fait partie de l'Institut Mines-Telecom. Elle est aussi chercheuse associée à l'Université Paris-Sud, RITM. Elle a soutenu sa thèse en économie à l'Université Paris-11 et à l'Université de Turin en Italie. Elle a été étudiante invitée à l'Université de Sussex à SPRU (2006-2007) et à ZEW à Mannheim (2013). Ses principaux intérêts de recherche sont reliés à l'économie des Technologies d'Information et Communication (TIC) et particulièrement à l'économie de l'innovation et des TIC, à l'économie de la vie privée et au *crowdfunding*. Elle a publié dans des journaux internationaux comme *Ecological Economics*, *Regional Studies*, *Research Policy*, *Telecommunications Policy*, *Technological Forecasting and Social Change*, *Industry and Innovation*.

Stephan CLÉMENÇON est professeur de mathématiques appliquées à TélécomParisTech, Institut Mines-Télécom, au sein du Département IDS (Image, Données, Signal) et anime le groupe de recherche S2A (Statistique, Signal et Apprentissage). Il effectue ses travaux de recherche en mathématiques appliquées au Laboratoire LTCI de Télécom ParisTech. Ses thématiques de recherche se situent principalement dans les domaines du *machine learning*, des probabilités et des statistiques. Il est responsable du Mastère Spécialisé « Big Data » à Télécom ParisTech et titulaire de la chaire industrielle « Machine-Learning for Big Data ».

Arnaud COUSTILLIÈRE, vice-amiral d'escadre, est né à Toulon le 3 novembre 1960. Sa carrière s'est essentiellement partagée entre des embarquements et commandements opérationnels sur des navires de combat et des postes de responsabilités en administration centrale, avec une spécialisation plus particulière pour les télécommunications, la cyberdéfense et la transformation. Le 1er septembre 2017 il est nommé directeur général des systèmes d'information et de communication du ministère des Armées, et élevé aux rang et appellation de vice-amiral d'escadre ; il est en particulier chargé d'orchestrer la transformation numérique du ministère. Il avait été nommé officier général à la cyberdéfense le 1er juillet 2011. Placé sous la double tutelle du chef d'état-major des armées et du chef de cabinet militaire du ministre, il était responsable de la montée en puissance de la cyberdéfense du ministère et de sa conduite opérationnelle. Depuis janvier 2017, il exerçait les fonctions de commandant de la cyberdéfense et disposait de l'état-major « COMCYBER » nouvellement créé. À l'état-major des armées depuis l'été 2008, il a été en poste d'officier de cohérence opérationnelle des armées en charge de la transformation du domaine des télécommunications et de la cyberdéfense. En 2006, il avait été nommé directeur des systèmes d'information de la marine dans un contexte de profonde réorganisation et mutualisation des systèmes d'information du ministère. Issu de la promotion 1981 de l'École navale, il a été embarqué sur de nombreux bâtiments de combat déployés en zones de crises, principalement en Méditerranée et dans le Nord de l'océan Indien. Il a exercé le commandement de la frégate lance-missiles *Duquesne*, des avisos *Commandant Bouan* et *D'Estienne d'Orves* ainsi que du bâtiment-école *Chacal*. Le vice-amiral d'escadre Arnaud Coustillière est officier de la Légion d'honneur, commandeur de l'ordre national du Mérite et commandeur de l'ordre de la Croix de l'Aigle de la république d'Estonie.

Éric FREYSSINET, colonel, est chef de la Mission numérique de la gendarmerie nationale depuis le 1er mai 2017, dans la continuité de vingt ans de carrière dans le domaine de la lutte contre la cybercriminalité à des postes techniques, stratégiques et opérationnels. Il a notamment exercé comme chef du département informatique-électronique de l'Institut de Recherche criminelle de la Gendarmerie nationale, chargé de projets Cybercriminalité à la sous-direction de la police judiciaire de la Gendarmerie nationale ou encore chef du Centre de lutte contre les Criminalités numériques. Dans ce cadre, il fut très impliqué dans la coopération internationale, notamment comme vice-président du groupe de travail d'Europol des chefs d'unités de lutte contre la cybercriminalité ou président du groupe de travail équivalent à Interpol. Plus récemment, il fut conseiller au sein de la Délégation chargée de la lutte contre les cybermenaces au ministère de l'Intérieur. Après une formation initiale d'ingénieur généraliste (École polytechnique, X92), le colonel Freyssinet s'est spécialisé dans la sécurité des systèmes d'information (Mastère spécialisé SSIR Télécom Paristech 99-2000) et a poursuivi dans une démarche par la recherche en soutenant une thèse de doctorat en informatique en 2015 (Université Pierre-et-Marie-Curie), sur le sujet de la lutte contre les *botnets*. Éric Freyssinet est membre associé du LORIA de Nancy.

Frédérick GARCIA, directeur de recherche à l'Inra, est membre de l'équipe « Modélisation des Agroécosystèmes et Décision » de l'unité MIA à Toulouse. Il est directeur-adjoint de l'institut Convergences Agriculture Numérique #DigitAg.

Yves GASSOT a été pendant plus de vingt ans directeur général de l'IDATE DigiWorld. À ce titre, il a participé à de très nombreux travaux sur l'évolution des marchés des télécommunications et plus largement du numérique, a assuré la direction de la revue *Communications* & *Stratégies,* du rapport annuel DigiWorld Yearbook, et il est l'auteur de nombreux articles. Il a été conseiller spécial de Viviane Reding, Commissaire européenne, pendant la revue du cadre réglementaire des communications électroniques. Il travaille aujourd'hui à la direction générale d'Orange. Yves Gassot est par ailleurs membre associé du Conseil général de l'Économie et du comité stratégique d'Iris Capital. Yves Gassot a une formation initiale d'architecte (DPLG-Paris) et une maîtrise d'urbanisme, et il est diplômé de l'IEP Paris (3e cycle d'aménagement).

Christine GATEAU est associée du département Contentieux du bureau de Paris du cabinet Hogan Lovells. Elle est considérée comme « *experte* » dans le contentieux relatif aux transactions en ligne ainsi que dans celui lié aux contrats IT et « *unique* » lorsqu'il s'agit de défendre des plateformes de commerce électronique. Reconnue pour son expertise dans le secteur des Technologies de l'Information ainsi qu'en matière de responsabilité du fait des produits, elle a développé une connaissance pointue des réglementations applicables au secteur des nouvelles technologies. Elle est également en charge de plusieurs dossiers actuellement pendants devant la Cour européenne de Justice ayant trait à des questions relatives à la protection des données personnelles – et préside par ailleurs le Comité international de l'association DRI – The Voice of the Defense Bar. Elle participe régulièrement en tant qu'intervenante à des séminaires et conférences en Europe, aux États-Unis et en Asie.

Pauline GIVORD est diplômée de l'École polytechnique et de l'Ensae et titulaire d'un doctorat et d'une habilitation à diriger les recherches en sciences économiques. Entrée à l'Insee en 1998, elle y a occupé plusieurs postes de chargée d'études économiques et elle a également été responsable de l'enquête Emploi. Après un passage par le Crest, centre de recherche de l'Insee, elle a dirigé de 2006 à 2012 une unité de chargés d'études sur les entreprises. Spécialisée dans l'utilisation des méthodes d'analyse quantitative d'évaluation des politiques publiques, elle a créé en 2012 une unité de statisticiens chargés de la diffusion des méthodologies statistiques innovantes. Elle est actuellement responsable de la transformation de cette unité en un SSPLab qui assurera la même mission au profit de l'ensemble du système statistique public.

Henri ISAAC, docteur en sciences de gestion, est maître de conférences à PSL Research University, Université Paris-Dauphine, et chercheur au sein de Dauphine Recherches en Management (CNRS, UMR 7088). Il a été directeur de la recherche et directeur académique à Neoma Business School (2009-2012) et vice-président « Transformation numérique » de l'Université Paris-Dauphine (2014-2016). Il dirige le Master Management Télécoms et Médias. Il est l'auteur de *E-commerce. De la stratégie à la mise en œuvre* (4e édition, 2017, Pearson France), et co-auteur de *Marketing digital* (2017, Pearson), *Travail à distance & e-management* (Dunod, 2013). Il est l'auteur de nombreux articles dans des revues académiques comme *Journal of Business Strategy*, *European Journal of Information Systems*, *International Journal of Innovation and Technology Management*, *International Journal of Mobile Communications*, *Revue française du Marketing*, *Système d'Information & Management*, *Revue française de Gestion*. Il est également président du *think tank* « Renaissance numérique ».

Laure KALTENBACH est cofondatrice et présidente de CreativeTech, entreprise dédiée aux relations entre arts, sciences et innovation technologique (www.thecreativetech.fr). Passionnée d'art contemporain et de neurosciences, elle est également fondatrice de Creative Futures. Elle est membre fondateur du laboratoire d'idées et des rencontres internationales du Forum d'Avignon

(www.forum-avignon.org) – culture, économie et innovation – dont elle a été directrice générale de 2007 à 2016. Elle a commencé sa carrière chez Accenture en média et télécommunications, où elle a passé plus de onze ans, puis deux ans chez TF1 et deux autres années dans les services du Premier ministre (DDM). Elle est membre du conseil d'administration de Cartooning for Peace, du HIP Institute et de l'Université d'Avignon. Elle intervient dans des conférences et dans des universités et enseigne dans le cadre du certificat MBA sur les industries culturelles et créatives de Panthéon-Assas. Elle publie régulièrement dans des revues et la presse. Elle est co-auteur aux Éditions First des *Nouvelles frontières du net, qui se cache derrière internet ?* (2010) et de *Paroles de Praticiens – Génération CreativeTech-ers* aux éditions Panthéon Assas, avec le professeur Jérôme Duval-Hamel, à paraître en 2018.

Matthieu MANANT est maître de conférences en sciences économiques à l'Université Paris-Sud et membre de l'Université Paris-Saclay. Il a soutenu sa thèse en économie à Télécom ParisTech en 2007. Ses centres d'intérêt sont l'économie de l'innovation et particulièrement les processus de coopération et de collaboration dans l'économie de télécommunication, et en économie numérique, avec une attention à l'économie de la vie privée, aux réseaux sociaux et à la régulation numérique.

Winston MAXWELL est avocat aux barreaux de New York et de Paris, associé du cabinet Hogan Lovells. Diplômé en droit (Cornell) et en économie (Telecom ParisTech), Winston Maxwell conseille des entreprises sur les contrats et sur la réglementation des données, des télécommunications, des médias et de l'Internet. Il codirige au niveau mondial la pratique « TMT » du cabinet Hogan Lovells. Il est membre du CSA Lab, un groupe de réflexion prospective réunissant des experts du numérique et de l'audiovisuel avec l'objectif d'anticiper et de caractériser les évolutions de l'économie et de la régulation audiovisuelles induites par la transformation numérique. En 2011, il a publié *La Neutralité d'Internet* avec Nicolas Curien (Éditions La Découverte), et en 2017 il a publié, dans la collection i³, l'ouvrage *Smart(er) Internet Regulation Through Cost-Benefit Analysis* (Presses des Mines), qui examine le problème de la régulation des contenus sur Internet.

Grégoire MOREAU est ingénieur diplômé de l'ENSEM (Nancy), promotion 1992. Il a effectué une première partie de carrière dans plusieurs fonctions au sein de la division Production nucléaire d'EDF. Depuis 2016, au sein de la direction Recherche et Développement du même groupe, il est responsable de l'équipe qui conçoit des solutions innovantes de surveillance des matériels des parcs de production d'électricité.

Pascal NEVEU, ingénieur de recherche à l'INRA et directeur de l'unité MISTEA, est membre du #DigitAg, où il coanime l'axe «systèmes d'information ».

Bertrand PAILHÈS est ingénieur (Télécom ParisTech) et diplômé de SciencesPo. Il travaille depuis une quinzaine d'années dans les administrations chargées du numérique en France (ARCEP, CNIL) et en cabinet ministériel entre 2013 et 2017. Il a été notamment directeur de cabinet d'Axelle Lemaire, secrétaire d'État au Numérique et à l'Innovation, et a coordonné la préparation et la mise en œuvre de la loi pour une République numérique, adoptée le 7 octobre 2016.

Anne PERROT a rejoint MAPP en 2012. Elle a été vice-présidente du Conseil, puis de l'Autorité de la Concurrence (2004-2012), et a participé à différents groupes d'experts dans le domaine de la concurrence (EAGCP auprès de la Commission européenne) et de la régulation auprès de régulateurs français (Commission Champsaur sur la dérégulation des télécoms, groupe d'experts de la CRE). Auparavant elle était professeur de sciences économiques à l'Université Paris-I et à l'ENSAE (1991-2004), directeur du Laboratoire d'Économie industrielle du CREST. Elle a publié dans de

nombreuses revues françaises et internationales dans les domaines de l'économie industrielle et de la politique de la concurrence. Elle a enseigné à l'École d'économie de Paris, à la Brussels School of Competition, et à Sciences-Po Paris. Elle est titulaire d'un doctorat de mathématiques, d'un doctorat de sciences économiques et est agrégée des universités en sciences économiques. En 2015, elle a présidé la commission d'évaluation de la loi pour la croissance, l'activité et l'égalité des chances économiques. Depuis 2015, elle est membre correspondant du Conseil d'Analyse économique, et depuis 2017 membre du Comité exécutif du Conseil national de l'Industrie.

Vincent PERTUY est architecte d'entreprise au périmètre des producteurs nucléaire et thermique chez EDF, il suit le développement des nouvelles technologies numériques.

Jean-Baptiste SOUFRON est avocat associé au cabinet FWPA Avocats. Après avoir travaillé sur les sujets de modélisation du droit à l'université Paris-2 Panthéon Assas, il a été le premier directeur juridique de la Wikimedia Foundation aux États-Unis. Il a créé et dirigé le *think tank* de Cap Digital. Il a ensuite été nommé conseiller à l'Économie numérique au cabinet du ministre en charge de l'Économie numérique, puis Secrétaire général du Conseil national du Numérique. Il est l'un des membres fondateurs du Hub France IA et publie et intervient régulièrement sur les questions du droit du numérique et des données.

Rémi STEINER est ingénieur général des Mines, actuaire et titulaire d'une maîtrise de droit des affaires. Il a exercé des responsabilités variées dans différents établissements bancaires, notamment en tant qu'administrateur et directeur général délégué des banques Hervet et UBP (Union de Banques à Paris), dont la fusion avec le CCF a donné lieu à la naissance de HSBC France. Il a rejoint depuis 2011 le Conseil général de l'Économie, dans le contexte où le champ d'expertise de cette entité, présidée par le ministre de l'Économie et des Finances, était étendu à l'ensemble des services financiers et aux activités qui s'y rattachent. Il est membre depuis 2012 du jury du concours d'adjoint de direction de la Banque de France.

Bruno SUTY a rejoint le Groupe EDF en 1987. Il est actuellement directeur des Systèmes d'Information et de la Transition numérique industrielle de la Direction du Parc Nucléaire et Thermique d'EDF.

Alexandre TERMIER, professeur d'informatique à l'Université de Rennes-1 et responsable de l'équipe-projet Inria/Irisa « Lacodam », est membre de #DigitAg, où il coanime l'axe « fouille de données ».

Patrick WAELBROECK est titulaire d'une thèse en économie de l'Université de Paris-1 Panthéon-Sorbonne et professeur d'économie industrielle et d'économétrie à Telecom ParisTech. Ses travaux portent sur l'économie de l'innovation, l'économie de la propriété intellectuelle, l'économie de l'Internet et des données personnelles. Il est membre du bureau de l'association EPIP (*European Policy for Intellectual Property*), dont il fut le président en 2013-2014. Il est cofondateur de la chaire Valeurs et politiques des informations personnelles de l'Institut Mines-Télécom qui aborde les problèmes des données personnelles et du Big Data sous différents angles : juridique, économique, technique et philosophique. Patrick Waelbroeck enseigne l'économie de l'Internet et des données dans le mastère spécialisé Big Data de Télécom ParisTech, ainsi qu'aux ingénieurs-élèves du corps des Mines. Il est *area editor* de la revue *Annals of Telecommunications* et membre du comité éditorial du *Journal of Cultural Economics*.