



Enjeux sociaux des Big Data

Paola Tubaro, Antonio Casilli

► **To cite this version:**

Paola Tubaro, Antonio Casilli. Enjeux sociaux des Big Data. Mokrane Bouzeghoub, Rémy Mosseri. Les Big Data à découvert, CNRS Editions, pp.292-293, 2017. <hal-01456369>

HAL Id: hal-01456369

<https://hal.archives-ouvertes.fr/hal-01456369>

Submitted on 4 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Enjeux sociaux des Big Data

Paola Tubaro et Antonio Casilli

Si les discours d'accompagnement des Big Data épousent souvent le point de vue des producteurs – que ce soit des organismes publics, des entreprises privées ou des plateformes numériques – il importe aussi de s'interroger sur le destinataire de leur action, l'utilisateur, qui joue tour à tour le rôle de citoyen, travailleur, consommateur et même objet d'observation ou « data point ». Que des données touchent directement les individus n'est pas un fait nouveau : une intense activité statistique publique, dont les origines remontent aux recensements de l'antiquité, aide déjà à la mise en œuvre de politiques au plus près de la vie de chacun, comme l'aménagement du territoire local ou l'octroi d'aides sociales. Bien évidemment, le revers de la médaille de cette mise en chiffres des populations est un effet de contrôle et de « gouvernement par les nombres », dénoncé notamment par Alain Desrosières (sociologue et statisticien à l'INSEE) en 2008. Historiquement, les retombées de l'usage de données sur le citoyen ont même pris des allures très sinistres, avec des pratiques de surveillance, de stigmatisation de parties de la population et de censure dans des régimes totalitaires. Ce qui change aujourd'hui avec les Big Data, est non seulement une augmentation substantielle de la masse de données collectées et traitées, grâce aux technologies numériques ; mais aussi, et surtout, l'entrée en force de l'entreprise privée en tant que productrice de données, devançant parfois les autorités publiques.

La promesse inhérente à ces transformations est celle d'une meilleure adéquation de l'offre de biens et services, publics ou privés, aux besoins qui émanent des personnes. Par exemple, des données sur le trafic, actualisées en temps réel et géolocalisées, permettraient une meilleure coordination des déplacements, et entraîneraient des gains de temps pour tous ainsi qu'une diminution des accidents. En santé, une connaissance plus fine de la diffusion de maladies, notamment contagieuses, perfectionnerait le dépistage précoce et la prévention (cf. I.9). Dans le contexte marchand, les données pourraient alimenter de nouvelles activités professionnelles, comme l'offre de services d'information ou conseil, susceptibles d'affiner la prise de décisions (cf. III.2). Certes, la réalisation concrète de ces promesses reste encore hypothétique dans de nombreux secteurs, mais le potentiel d'amélioration des conditions de vie de chacun est important, et les progrès sont incessants.

Menace sur la vie privée

Pour autant, ces promesses s'accompagnent de coûts, en partie déjà observés, en partie anticipés, dont il convient de tenir compte afin de prévoir des pistes de développement qui respectent la

dignité et les libertés fondamentales des individus, et concilient développement technologique et bien-être des personnes. Un premier enjeu est la protection de la vie privée, menacée malgré les lois de protection des données personnelles déjà en vigueur dans plusieurs pays européens, et les traités qui régissent leur circulation au niveau international (cf. VIII.5). Toute collecte et tout traitement de Big Data dans le contexte socio-économique actuel impliquent une récupération massive d'informations sur les individus, comme leurs déplacements (validations dans les transports en communs, péages d'autoroute), leurs dépenses (logs d'activité des cartes bancaires ou des cartes de fidélité), leur santé (prestations délivrées par le biais d'une carte Vitale), leurs opinions et comportements (blogs, Twitter, Facebook)... La facilité accrue de croiser ces données pour dessiner un portrait fin de chaque individu est porteuse de risques de dérives, de la part d'acteurs publics tout autant que privés. La surveillance étatique sur les citoyens est redevenue d'actualité dans les pays occidentaux avec les scandales qui ont suivi les révélations d'Edward Snowden en 2013. L'intrusion des entreprises n'est pas moins à craindre, pouvant déboucher sur des formes de discrimination par ailleurs interdites, pénalisant éventuellement les individus pour leur genre, âge, origine ethnique, lieu de résidence ou niveau socio-économique.

Exploitation des usagers

Un deuxième enjeu est le travail souvent invisible et non rémunéré qui alimente les bases de données et les algorithmes qui permettent d'en extraire de l'information, par exemple de reconnaître des images ou classer des documents. Certaines de ces opérations sont très difficiles à automatiser et les algorithmes doivent être « entraînés » sur la base d'un corpus d'exemples pertinents : ceux-ci sont fournis par les usagers des plateformes numériques ou applications qui, eux, savent trier, qualifier et filtrer des contenus. Des services comme Amazon Mechanical Turk ou Google Crowdsourcing représentent des exemples de l'implication constante de foules d'êtres humains dans la réalisation de tâches de calibrage de logiciels de traduction, de reconnaissance, d'étiquetage. Un véritable marché du travail du « calcul à base d'humain » sous-tend les Big Data et les algorithmes qui les traitent. C'est ce que l'on appelle le *digital labor* : des microtâches à faible intensité, réparties entre des milliers d'usagers et souvent non reconnues comme travail, mais qui aident les intelligences artificielles encore peu évoluées, et dont les grandes entreprises du Web extraient de la valeur. Par exemple le Re-Captcha (cf. V.11), à l'apparence un instrument d'authentification distinguant un usager humain d'un robot, est aussi utilisé pour améliorer la reconnaissance de nombres et de mots, afin d'affiner les méthodes de numérisation de sources documentaires ou d'images, mobilisables à des fins commerciales. Le problème de la répartition de la richesse produite par ces activités se pose, l'« exploitation » actuelle des usagers apparaissant de moins en moins soutenable – surtout face à la montée de recours collectifs en justice et de mouvements organisés

revendiquant la reconnaissance de protection sociale pour les « travailleurs du clic » et l'accès et le contrôle des données et de leur valorisation.

Opacité sur le partage des données

Un troisième ensemble de risques concerne la transparence. Contrairement aux données « classiques » de la statistique publique, les Big Data sont souvent la propriété privée des entreprises, des plateformes et des courtiers en données (*data brokers*). Leur partage n'est guère la règle, car même lorsqu'elles sont mises à disposition, les Big Data sont peu lisibles et peu de producteurs sont prêts à payer le coût élevé de leur documentation et curation. Surtout, le manque de transparence concerne les algorithmes d'extraction et traitement des données, dont le fonctionnement n'est presque jamais explicité, pour des raisons qui relèvent de leur haute complexité technique, ainsi que du secret commercial. Les indicateurs publiés peuvent alors apparaître opaques, et il devient difficile pour l'individu non initié de savoir quelle confiance leur accorder. En l'absence d'un « audit citoyen des algorithmes », l'utilité publique des Big Data est mise en cause.

Pour conclure, il devient nécessaire de réfléchir à une régulation adaptée pour que les Big Data rapportent à l'utilisateur plus qu'elles ne lui coûtent. Sans entraver les avancées techniques et les opportunités économiques en résultant, il faut tenir compte des risques inhérents aux données numériques, en termes de protection des droits et de l'information personnelle des individus.

Bibliographie

Bénabou V.L. et Rochfeld J. (2014). *A qui profite le clic ?* Editions Odile Jacob.

Cardon D. & Casilli A.A. (2015). *Qu'est-ce que le Digital Labor ?* INA éditions.

Casilli A.A. (2014). « Quatre thèses sur la surveillance numérique de masse et la négociation de la vie privée ». In J. Richard et L. Cytermann, *Etude annuelle 2014 du Conseil d'Etat « Le numérique et les droits fondamentaux »*, La Documentation Française, pp.423-434.

Desrosières A. (2008). *Gouverner par les nombres : L'Argument statistique II*, Presses de l'École des Mines de Paris.

Affiliation

Antonio A. CASILLI. Sociologue, Maître de conférences à Télécom-ParisTech, Laboratoire i3, Paris.

casilli@telecom-paristech.fr

Paola TUBARO. Sociologue et économiste, Chargée de Recherche au CNRS, Laboratoire de Recherche en Informatique, Orsay. paola.tubaro@lri.fr

Glossaire :

Digital labor : tâches réalisées par les usagers des plateformes numériques, qui sont assimilables au travail dans la mesure où elles produisent de la valeur économique, mais qui ne font pas nécessairement l'objet d'un contrat ou d'une rémunération.

Data broker : en français courtier en données. Il s'agit d'un intermédiaire (individu ou firme) pratiquant le commerce de données relative aux internautes, à l'intention d'entreprises voulant peaufiner leur ciblage publicitaire.

Data curation : l'ensemble des activités finalisées à l'archivage, la conservation, la documentation et la mise à disposition des données.