



HAL
open science

Big data ou l'illusion d'une synthèse par agrégation. Une critique épistémologique, éthique et politique

Léo Coutellec, Paul-Loup Weil-Dubuc

► To cite this version:

Léo Coutellec, Paul-Loup Weil-Dubuc. Big data ou l'illusion d'une synthèse par agrégation. Une critique épistémologique, éthique et politique. Journal international de bioéthique et d'éthique des sciences , 2017, Données de masse en santé, 28, pp.63-79. <10.3917/jib.283.0063>. <hal-03347985>

HAL Id: hal-03347985

<https://hal.science/hal-03347985v1>

Submitted on 18 Sep 2021

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.



HAL Authorization

CHAPITRE 7. *BIG DATA* OU L'ILLUSION D'UNE SYNTHÈSE PAR AGRÉGATION. UNE CRITIQUE ÉPISTÉMOLOGIQUE, ÉTHIQUE ET POLITIQUE

Léo Coutellec, Paul-Loup Weil-Dubuc

ESKA | « *Journal International de Bioéthique* »

2017/3 Vol. 28 | pages 63 à 79

ISSN 2555-5111

ISBN 9782747227254

Article disponible en ligne à l'adresse :

<https://www.cairn.info/revue-journal-international-de-bioethique-2017-3-page-63.htm>

Pour citer cet article :

Léo Coutellec, Paul-Loup Weil-Dubuc « Chapitre 7. *Big data* ou l'illusion d'une synthèse par agrégation. Une critique épistémologique, éthique et politique », *Journal International de Bioéthique* 2017/3 (Vol. 28), p. 63-79.
DOI 10.3917/jib.283.0063

Distribution électronique Cairn.info pour ESKA.

© ESKA. Tous droits réservés pour tous pays.

La reproduction ou représentation de cet article, notamment par photocopie, n'est autorisée que dans les limites des conditions générales d'utilisation du site ou, le cas échéant, des conditions générales de la licence souscrite par votre établissement. Toute autre reproduction ou représentation, en tout ou partie, sous quelque forme et de quelque manière que ce soit, est interdite sauf accord préalable et écrit de l'éditeur, en dehors des cas prévus par la législation en vigueur en France. Il est précisé que son stockage dans une base de données est également interdit.

BIG DATA OU L'ILLUSION D'UNE SYNTHÈSE PAR AGRÉGATION. UNE CRITIQUE ÉPISTÉMOLOGIQUE, ÉTHIQUE ET POLITIQUE

*L. COUTELLE*¹ & P.-L. WEIL-DUBUC**

INTRODUCTION : L'AGRÉATION COMME NŒUD PROBLÉMATIQUE DES *BIG DATA*

« On fait la science avec des faits comme on fait une maison avec des pierres : mais une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison. » (Henri Poincaré)

Tendanciellement les données, et en particulier les données de santé, deviennent constitutives de notre environnement le plus immédiat, un peu comme l'air

* Département de recherche en éthique, Université Paris-Sud Paris-Saclay, Espace éthique Ile-de-France, Laboratoire d'excellence Distalz, France.

¹ leo.coutellec@u-psud.fr

que nous respirons : nous en produisons et nous y sommes confrontés sans y prendre garde. Il y a désormais une apparente immédiateté dans notre rapport aux données. Par exemple, si l'accès aux données de santé s'est partiellement affranchi de la médiation du médecin, il s'est aussi, dans le même temps, apparemment affranchi de la médiation de la personne concernée elle-même qui, sans choix, sans demande, peut se voir confrontée à la révélation de ses propres données. Cette pseudo-immédiateté de notre rapport aux données s'accompagne d'une tendance à les mystifier et, plus précisément, à conférer à ces données ainsi immédiatement produites, disponibles et susceptibles de s'imposer à nous des propriétés magiques. Les données pourraient dire tout à la fois la totalité du monde, du moi et de ce qui fait société. Elles auraient non seulement vocation à dire ce qui est mais aussi ce qui adviendra. Ces croyances sont constitutives du phénomène *big data* que nous devons ainsi appréhender comme enjeu à la fois culturel, scientifique et technique (Boyd & Crawford, 2012) ; elles nous obligent à investir un terrain d'analyse où se rencontrent épistémologie, éthique et politique².

Dans ce papier, nous proposons une approche critique du phénomène *big data* par la déconstruction du principe méthodologique qui en structure la logique : le principe d'agrégation. Notre hypothèse se situe ainsi en amont, ou en deçà, des critiques qui font de l'usage d'outils de collecte et de traitement automatique de données massives et hétérogènes un nouveau mode de gouvernement : gouverner par les standards (Frydman & Van Waeyenberge, 2013), gouverner par les nombres (Desrosières, 2008 ; Supiot, 2015) ou gouverner par les algorithmes (Rouvroy & Stiegler, 2015), variations d'une nouvelle forme de gouvernementalité néolibérale (Dardot & Laval, 2009). Parce que concentrées sur les dispositifs (algorithmes, standards, protocoles) selon une approche quasi-phénoménologique, ces critiques doivent être complétées par une analyse épistémologique qui sous-tend la possibilité même de leur existence. L'agrégation, comme mode de traitement de l'hétérogénéité des données, structure la pensée *big data*, elle en est la logique même, l'agrégat en est l'objet. D'un point de vue épistémologique, le noyau des *big data* peut se formuler de la façon suivante : fragmenter pour mieux agréger, agréger pour mieux fragmenter. Cette dialectique fragmentation/agrégation repose sur une présomption d'agrégabilité généralisée (tout peut s'amalgamer au sein du milieu technique des données, la base de données)

² Cette réflexion théorique s'appuie sur un terrain pratique au sein du laboratoire d'excellence Distalz qui vise à mieux comprendre les mécanismes physiopathologiques impliqués dans le développement de la maladie d'Alzheimer et à explorer les mutations épistémologiques, éthiques et socio-politiques induites par la possibilité d'un diagnostic précoce de cette maladie. L'utilisation de données massives et hétérogènes (génomiques, biologiques, radiologiques, cliniques) joue un grand rôle dans ces recherches ; la perspective d'un diagnostic pré-symptomatique ou asymptomatique de la maladie d'Alzheimer impose de réfléchir sur les apports de la démarche *big data*.

et sur la prétention de faire de l'agrégation la voie privilégiée pour la production de nouvelles synthèses.

Nous procédons en trois temps pour déconstruire cette idée et défaire la prétention de l'agrégation à s'imposer comme nouvelle voie pour produire des connaissances (critique épistémologique), comme nouvelle synthèse identitaire (critique éthique) et enfin comme nouveau modèle de solidarité (critique politique). Chaque fois nous montrerons que ces tentatives d'agrégation échouent à produire leurs objets : ni connaissance, ni identité, ni solidarité ne peuvent résulter d'un processus d'amalgame. Dans les trois cas, l'agrégation s'accompagne toujours d'un moment de fragmentation dont la dissociation, la dislocation et la séparation sont différentes figures. Ce travail doit nous permettre d'envisager autrement la critique des *big data* en déplaçant le regard des usages vers l'essence méthodologique d'une telle démarche. Le pari que nous faisons alors est de faire hésiter ce qui se présente comme une nouvelle façon de penser l'homme et le monde (Babinet, 2015).

1. LA PRODUCTION D'UNE CONNAISSANCE COMME AGRÉGAT

Parmi les prétentions qui président à la mise en œuvre d'une démarche *big data*, celle de contribuer à une nouvelle forme de production des connaissances, une nouvelle ère pour la méthode scientifique anime les discussions épistémologiques contemporaines. La *science-data-driven* marquerait l'avènement d'un quatrième paradigme après la *science empirique* décrivant les phénomènes, la *science théorique* usant des modèles et des généralisations, et la *science computationnelle* simulant les phénomènes complexes (Hey *et al.*, 2009)³. L'une de ses caractéristiques serait de substituer à la théorie ou à l'hypothèse une démarche plus agnostique (sans *apriori* théorique), faisant de l'immanence des données l'ultime source de la connaissance et de l'agrégation son principe méthodologique. Plus précisément, les caractéristiques épistémologiques principales de ce qui s'affiche comme nouveau paradigme sont les suivantes : la fin du modèle hypothético-déductif au profit d'une démarche empirico-inductive où ce sont les données qui guident l'exploration cognitive et non plus l'hypothèse ou la théorie (Anderson, 2008) ; l'exhaustivité remplace l'échantillonnage pour une collecte systématique et parfois aveugle de données (pensons ici au séquençage très haut débit du génome où il devient possible de « tester » les 20 000 gènes d'une cel-

³ Nous n'abordons pas ici les nombreuses apories qui accompagnent la faiblesse de ce genre de découpage historique, à l'instar de la théorie des mode1/mode2 de la science. A ce propos, voir : (Lamy, 2007).

lule en même temps et sans ciblage) ; l'apparition de corrélations remplacent la recherche de causalités au profit d'une rationalité statistique bien particulière « dès lors qu'elle tire sa force du traitement automatisé d'informations dont la seule qualité est l'aspect massif : indifférente aux causes des phénomènes, cette rationalité s'ancre dans l'observation purement statistique de corrélations (indépendantes de toute logique) entre données recueillies d'une manière absolument non sélective dans une variété de contextes hétérogènes les uns aux autres » (Rouvroy, 2010). Se dessine ainsi la volonté de produire une connaissance par agrégation de données massives et hétérogènes, une connaissance comme agrégat. Nous pouvons ici faire au moins trois critiques d'ordre épistémologique pour déconstruire cette prétention de la *science-data-driven* à produire une connaissance par agrégation.

1.1. Les données ne sont pas données : l'agrégat comme fiction

La possibilité de construire une connaissance comme agrégat – c'est-à-dire comme résultante d'un processus d'agrégation – tient d'abord à une propriété que l'on accorde à la donnée alors comprise comme *data*. Il s'agit de la capacité à pouvoir apparaître de soi-même, comme la réalité elle-même, à se présenter comme de purs « donnés » (Coutellec, 2015). Lorsque par exemple nous évoquons la découverte « fortuite » d'une donnée de santé potentiellement préoccupante (cela devient fréquent avec l'imagerie médicale), nous cédon précisement à l'illusion selon laquelle la donnée surgirait de l'inconnu, d'elle-même, oubliant alors les dispositifs techniques de production des données visant précisement à faire émerger ces données. Si la découverte est au mieux « inattendue », au moins pour le médecin et le patient, elle ne doit rien au hasard. Autrement dit, s'impose l'exigence de suivre précisement l'histoire phénoménotéchnique⁴ de la donnée, de son statut de *data* à celui de *ficta*. De la donnée brute (*data*) à la connaissance en passant par l'information pertinente, c'est tout un itinéraire technique, scientifique et social à reconstruire. Une telle démarche ne peut se faire qu'à la condition d'éviter deux illusions épistémologiques : (i) la neutralité des dispositifs de collecte et de traitement des données qui pourrait nous faire croire à une immanence radicale de la donnée qui par elle-même et par ses agrégations en grand nombre serait porteuse de sens ; (ii) la fin du modèle hypothético-déductif qui laisserait place à une démarche scientifique agnostique où l'analyse des données se ferait sans hypothèses préalables et sans modèles. Ces deux biais se renforcent dans les discours pour faire de la *science-data-driven* (ou la science axée sur les

⁴ Bachelard nous invitait à étudier les phénomènes dans la précision de leur technicité et dans l'enchevêtrement qu'ils impliquent entre théorie et pratique, il en appelait ainsi à une phénoménotéchnique. Voir : (Bachelard, 1965 ; pp. 9-10).

données), le nouveau *mainstream* scientifique qui remplacerait les précédents. L'emballement est beaucoup plus réservé dans le champ académique où cette « révolution des données » est plutôt ré-encadrée dans un ensemble plus vaste de démarches dans lesquelles hypothèses, modèles et théories gardent toute leur place. Cette prudence épistémologique face aux *big data*, que l'on retrouve notamment dans le champ de la recherche épidémiologique et génétique, tient à une caractérisation nuancée de ce l'on appelle « donnée ». Le postulat est double : *une donnée ne nous est jamais donnée*, elle est prise dans un dispositif de collecte dont le paramétrage et le calibrage (la métrologie) dépendent de certaines hypothèses scientifiques et d'intentions de recherche; une donnée n'est plus une *data* à partir du moment où l'on applique sur elle un dispositif technique et une série de filtres interprétatifs (Bollier, 2010), ce dont nous avons à faire est donc plutôt de l'ordre du *big ficta*. A cela se rajoutent des considérations strictement méthodologiques sur la qualité des données et la capacité d'agrégation de leur hétérogénéité qui dépend de la qualité du processus de traitement de l'information et de l'harmonisation des moyens de collecte. Pour toutes ces raisons, la connaissance comme agrégat ou résultat d'une agrégation n'est que fiction, et doit être traité comme telle au risque de naturaliser ce qui n'est qu'une opération de synthèse.

1.2. Le gap entre capacités interprétatives et capacités computationnelles : l'agrégat comme dissociation

La démarche *big data* a aussi pour conséquence d'éloigner la capacité de collecte de la capacité d'analyse ou, autrement dit, l'agrégat de données qui en est issu creuse l'écart entre capacité computationnelle et capacité interprétative. Ce phénomène de dissociation prend deux formes : une dissociation de sens (disruption) et une dissociation temporelle (désynchronisation). L'amalgame de données, agrégées au sein d'une base de données, ne produit pas tant de la connaissance que de l'information, et elle en produit dans des quantités devenues gigantesques. L'entreprise de sens qui permet le passage de l'information à la connaissance se retrouve ainsi contraint par le fait même que l'agrégat porte en lui-même un sens, immanent aux règles statistiques de corrélation qui lui donne sa forme au sein des bases de données. Avec la connaissance-agrégat, le sens comme exercice de création de signification ou de symbolisation n'est plus ce qui enrichie les données par des apports théoriques, historiques ou encore contextuels mais ce qui *prend forme* au sein d'un ensemble statistique. La connaissance comme agrégat n'est en fait qu'un amalgame d'informations auto-signifiantes. Avec l'avalanche de données, et son corollaire la fabrication d'agrégats cognitifs, nous assistons bien à un processus de dissociation. Le milieu associé (Simondon, 2012) aux agrégats de données n'est pas un milieu techno-

symbolique, au sens où il ne contribue pas à une entreprise de sens, mais un milieu numérique forclos. L'agrégat de données issues du séquençage très haut débit d'un génome humain n'est pas une connaissance mais un ensemble d'informations dont l'analyse et l'interprétation sont lourdement contraintes par l'existence même de cet agrégat massif. Le sens numérique des agrégats, de par sa puissance, sa rapidité et son caractère massif inhibe les tentatives interprétatives qui viseraient à contextualiser, à théoriser, à conceptualiser l'agrégat. Ce dernier, constitutivement, résiste ainsi à toute tentative interprétative qui ne soit pas elle-même sous l'emprise de la base donnée. Cette *data society* construite non plus sur des connaissances signifiantes mais sur des agrégats de données contribue au caractère disruptif de nos sociétés (Stiegler, 2015). Nous le voyons, la démarche *big data*, lorsqu'elle prétend à produire une connaissance comme agrégat, devient hermétique à toute critique.

La dissociation produite par l'agrégation massive de données prend aussi la forme d'une désynchronisation. L'agrégat de données permis par les démarches *big data* a cette caractéristique d'accompagner un processus sans précédent d'accélération du rythme de collecte et de production de base de données. L'accélérateur européen de particules produit chaque jour 40 000 Go de données ; la bibliothèque médicale *PubMed* contient plus de 22 millions de publications accessibles gratuitement en ligne, et 40 000 à 50 000 de plus viennent s'y ajouter chaque mois ; *Facebook* traite en moyenne 50 milliards de photos ; un cerveau analysé par IRM génère en quelques secondes 8 Go de données. Séquencer le génome complet de 1 000 individus peut se faire en quelques jours et nécessite des capacités de stockage de l'ordre de 500 To. Se produit inévitablement une désynchronisation majeure entre ce rythme et cette vitesse de génération de la donnée et le rythme ou la temporalité de son appropriation. Ainsi s'impose la norme temporelle silencieuse de l'accélération (Rosa, 2012) qui propulse l'agrégat de données dans toutes les sphères de nos existences, inhibant toute capacité de pause temporelle ou de décélération qui pourrait être favorable à une appropriation collective de ces agrégats pour en faire des connaissances pertinentes. La compréhension du monde et les explications qui en résultent habitent désormais l'environnement temporelle des *data center*.

1.3. Le traitement de l'hétérogénéité forte : l'agrégat comme standard

La troisième critique que nous pouvons faire à la prétention de produire une connaissance-agrégat concerne le traitement de l'hétérogénéité forte des données⁵. Données génétiques, biologiques, de santé, d'imagerie, liées aux modes de

⁵ Par « hétérogénéité forte », nous entendons l'hétérogénéité des mondes et non plus seulement celle des disciplines.

vie ou aux habitudes alimentaires pourraient ainsi s'agréger au sein d'un même ensemble, au sein d'une même *base* de données pour établir des corrélations jusqu'alors inaccessibles dans le cadre d'analyse en silo ou en groupe thématiques de données (démarche classique de l'épidémiologie). Nous partageons cette intention de décloisonnement et de transversalité, cette idée que la complexité du réel doit se refléter dans la complexité des outils pour l'appréhender. Mais, paradoxalement, ce traitement de l'hétérogénéité forte par l'agrégation ne se fait que sous la contrainte d'une forme de standardisation : la définition d'une convention d'équivalence (numérique) par la présomption d'agrégabilité selon laquelle il n'y a pas d'incompatibilité ou d'incommensurabilité indépassable entre type et source de données. La convention d'équivalence généralisée des *big data* est la donnée numérique, une donnée à laquelle nous donnons une forme numérique pour qu'elle devienne une information que l'on peut traiter au sein de base de données. Il s'agit là d'une forme de standardisation via l'établissement d'une convention d'équivalence qui agrège des réalités différentes en présumant leur agrégabilité numérique. Cette critique des conventions d'équivalence comme processus de standardisation résulte de la façon dont on conçoit l'activité de quantification. Si *quantifier*, c'est « exprimer par des nombres ce qui l'a été auparavant par des mots », il faut aussi rappeler que la quantification se distribue toujours en deux moments : « se mettre d'accord sur des conventions, des choix d'objets et de mises en équivalence, puis, une fois arrêtées ces conventions, de procéder aux opérations de mesure proprement dites » (Desrosières, 2014, p. 236). Les démarches *big data* conviennent que tout peut s'agréger numériquement, autrement dit que la base de donnée relève d'une forme d'inclusivisme⁶, à partir du moment où l'on a accepté de traiter l'hétérogénéité forte sous le mode de sa standardisation numérique. L'agrégat, ainsi résultant, devient un nouveau standard là où la prétention de départ était d'en faire un objet complexe.

La synthèse numérique qui vise à produire une connaissance-agrégat entend aussi s'imposer comme synthèse identitaire.

2. LA RECHERCHE D'UNE SYNTHÈSE IDENTITAIRE PAR AGRÉGATION

Affirmer, sans préjuger de ce que nous sommes, que nous ne sommes pas nos données est une banalité. Nous le sentons bien en nous-mêmes et aucun individu sensé ne pourrait sérieusement adhérer à cette croyance que ses données sont des

⁶ L'*inclusivisme* est un concept théologique qui désigne une volonté de voir en chacun un chrétien implicite. Nous l'utilisons ici dans un sens épistémologique pour qualifier cette posture qui consiste à inclure dans un même ensemble des éléments hétérogènes.

parties de lui-même ou encore qu'une personne est réductible à ce que nous nommons, sur les réseaux sociaux notamment, un « profil ». Toutefois, il reste toujours possible de croire que l'agrégation de nos données ne fournit, en l'état actuel des connaissances scientifiques, que des approximations mais que la synthèse de la personne humaine par agrégation de données demeure un horizon accessible à l'intelligence humaine. C'est ce projet d'une numérisation des existences que défendent certaines théories transhumanistes visant l'immortalité : l'individu pourrait renaître de ses données, lesquelles pourraient s'apparenter à ses atomes constitutifs.

2.1. Le moi comme agrégat de données ?

La position atomiste, selon laquelle l'individu résulte d'une agrégation d'éléments constitutifs ou atomes, n'est pas neuve. Déjà présente chez les Épicuriens, elle s'est notamment vue réinvestie dans l'histoire récente de la génomique à la faveur des hypothèses généticistes aujourd'hui mises en cause, selon lesquelles l'identité et le devenir de l'individu pourraient se concevoir à partir d'une connaissance fine de la combinaison unique de ses gènes (Fox Keller, 2003 ; Darrason, 2014). Cette perspective physicaliste est aussi celle des partisans d'une approche neuroscientifique de l'individu. L'analyse et la critique de cette position dépassent largement le cadre du présent article : il nous importe ici de voir en quoi les *données* ne sauraient en aucun cas tenir lieu de ces atomes constitutifs de la personne et, plus encore, en quoi « un atomisme de la donnée » est une contradiction dans les termes.

La donnée n'est pas, comme le serait un atome, un élément premier qui nous serait donné. Nous pouvons citer au moins trois traits qui distinguent la donnée de l'atome. En premier lieu, la donnée est composite, parce qu'elle est le produit d'une addition entre éléments ou le résultat d'une succession d'événements (par exemple le nombre de pas). En second lieu, une donnée n'a pas d'existence propre puisqu'elle ne peut être appréhendée indépendamment de la présupposition d'autres données. Supposons par exemple que l'on trouve chez un être humain et un cochon un taux identique d'hypertension artérielle. Il serait absurde de dire qu'ils ont en partage *une même donnée*, comme nous pourrions dire qu'ils ont un gène commun. Une donnée ne peut exister comme donnée qu'au sein d'une base de données dont on a préalablement délimité le territoire et exclu des groupes d'individus, eux-mêmes sources de données potentielles. Comme nous l'avons écrit précédemment, une base de données se constitue toujours à partir d'une présomption d'agrégabilité entre données hétérogènes. Enfin, la métaphore de l'atome est trompeuse puisqu'elle tend à présenter le profil comme une simple agrégation de données préexistantes. Or, le processus de formation d'un profil est plus complexe puisqu'il implique en même temps qu'une opération d'agrégation

une *opération de dislocation*. À mesure que s'agrègent les données, le profil se détache de la population à laquelle il appartient, en appartenant à des groupes d'individus de plus en plus restreints jusqu'à se particulariser en entité unique.

Ce dernier point est d'autant plus important à garder à l'esprit que la possibilité d'un profilage entretient le projet d'une meilleure prise en compte des besoins de la personne par des offres dites « personnalisées ». Cette tentation concerne de nombreuses activités depuis la médecine jusqu'à l'architecture en passant par la coiffure ou l'assurance. Nos données dites « personnelles » ne sont pourtant que des éléments distinctifs mis en évidence par dislocation. L'identification ainsi obtenue s'effectue de l'extérieur et à partir d'un ensemble plus large (la population), comme du reste s'effectue tout état civil ; elle n'est pas le résultat d'une subjectivation, d'une reconnaissance de soi par soi, ni même le résultat d'un assemblage unique d'éléments premiers. En aucun cas donc, un profil ne peut ni pourra être assimilé au soi. Le profil est une représentation appauvrie de l'humain, un « homme disloqué » (Grimaldi, 2001 ; Benasayag, 2016), puis agrégé.

2.2. Le moi comme agrégateur dynamique de données ?

Le projet d'une création de soi par agrégation de données peut toutefois prendre une autre voie lorsque le moi est moins appréhendé comme le résultat d'une agrégation de données que comme l'agent dynamique de cette agrégation. Cette hypothèse nous conduit à envisager une autre figure de la notion de « profil » qui n'est plus alors un ensemble de données automatiquement générées et agrégées par des algorithmes, mais un sujet qui se construit en s'agrégeant dans un rapport dynamique et critique à l'égard de ses données personnelles. Dans cette perspective, je suis et je ne suis pas mes données, puisque leur connaissance me donne précisément l'opportunité de ne plus les subir comme des données mais de m'adapter à elles ou éventuellement de les modifier. Le profil devient alors l'objet d'une réagrégation constante de soi par soi. L'expression « *quantified self* » est également utilisée pour ce que nous nommerons ici « l'agent-profil ».

Le présupposé est le suivant : l'individu des sociétés contemporaines gagnerait, par l'accès à ces données et en particulier à ses données, non seulement en capacité à choisir son projet de vie, mais plus encore à réaliser les possibilités de son existence. Cette idée s'illustre dans le domaine de la santé par une « rhétorique de l'empowerment » (Juengst *et al.*, 2012) relayée par les firmes de commercialisation de tests génétiques, par les instituts de recherche médicale et les autorités publiques. La firme DeCODEme assurait que la « médecine prédictive » fournit une « carte routière pour améliorer sa santé » (Decodeme.com, 2009). La métaphore parle d'elle-même : le décodage du génome ferait apparaître divers *possibles* que l'individu pourrait actionner selon son projet de vie. Fort de cette

visibilité de ses données, il pourrait élaborer son existence selon un plan correspondant à ses conceptions de la vie bonne, de moins en moins tributaire des aléas et des agents extérieurs. Cette maîtrise serait ainsi directement proportionnelle à l'acquisition de ces données : plus l'individu connaît ses risques, plus il est en mesure d'établir une cartographie complexe de son existence, plus il la maîtrise de part en part, plus il se réalise dans l'actualisation volontaire de ses propres données.

Il est intéressant et important de remarquer que cet idéal de l'auto-réalisation par la connaissance de ses données correspond étroitement à deux thèses philosophiques que l'on peut toutes deux relier à l'idéologie libérale. La première consiste à penser la dignité de l'être humain comme capacité à s'arracher à la nature en lui-même, autrement dit à se penser indépendamment de l'ensemble de ses appartenances et de ses déterminations. Le parallèle est alors aisé à établir : l'agent-profil est une figure de cet individu libéral, en ce qu'il est capable de penser son existence tout à la fois comme indépendante des données qui le caractérisent et toujours cependant en relation avec elles. La seconde thèse philosophique renvoie au concept d'« individualité » développé par J. S. Mill (Mill, 1990). L'individualité est un processus dynamique de développement de capacités innées et acquises, et non une entité fixe et définie. L'idée de Mill est reprise plus récemment par J. Rawls dans *Théorie de la justice* (Rawls, 1987) à travers son concept de « plan de vie », système dynamique de fins hiérarchisées et organisées dans le temps qui favorise le développement de soi et octroie à l'être humain sa dignité de « personne » (voir aussi Audard, 2016).

Nous ne souhaitons pas, dans le présent article, prolonger la discussion philosophique déjà riche autour de ces thèses libérales. Il nous importe plutôt de voir ici sous quelles conditions théoriques la rhétorique de l'agent-profil décrite plus haut reprend à son compte cette idéologie libérale. Plus clairement dit, il s'agit pour nous de dénouer deux simplifications théoriques par lesquelles cette rhétorique parvient à revendiquer l'héritage de l'idéologie libérale.

La première simplification tend à assimiler les données à des *dons* que l'individu serait en mesure de cultiver ou encore à des capacités. Les données ne révèlent en effet aucune capacité (au sens de *power*) biologique sommeillant en l'être humain et qu'il pourrait actualiser, mais seulement des *probabilités*. Or les probabilités se distinguent des capacités par le fait qu'elles ne sont pas « actualisables ». En premier lieu, elles n'indiquent pas un *telos*, un bien à atteindre, mais le plus souvent un mal ; en second lieu, les probabilités ne dépendent pas ou pas seulement, pour leur probable réalisation, des efforts de l'agent. Les données génomiques, par exemple, peuvent nous dire que nous pouvons fumer sans risque particulier ; en aucun cas, elles ne nous disent que nous sommes particulièrement résistants aux effets du tabac et de quelles façons nous pourrions actualiser ou cultiver cette capacité de résistance au tabac. Tout au plus les données biolo-

giques constituent-elles des outils pour aménager son existence dans un champ de contraintes et de permissions prédéfini. L' « autonomie » ainsi promue, de façon trompeuse, comme *empowerment* revient en réalité à une forme de capacité à préempter un futur probable ; elle équivaut à une forme de prudence par laquelle celui qui connaît ses données personnelles peut ajuster son existence présente pour s'éviter d'être pris au dépourvu, en atténuant la probabilité d'un scénario non-souhaité et/ou le choc suscité par la réalisation éventuelle de ce scénario. Cet *empowerment* revient au mieux à donner le sentiment d'une maîtrise de l'existence – le sentiment que l'on agit plutôt dans le bon sens, sans en être certain ; au pire, à donner l'illusion de cette maîtrise – par le lissage émotionnel de la surprise face à l'incertitude tout au long de l'existence.

La seconde simplification tend à confondre conscience de ses propres déterminations, telles qu'elles se formalisent dans les données, et capacité à agir sur ces dernières ; pour le dire dans des termes anglais ici plus explicites, il convient de bien distinguer *self-awareness* et *self-control*. Dans *Le Point de vue de nulle part* (Nagel, 1993), le philosophe Thomas Nagel pointe l'illusion d'un élargissement de la liberté par la conscience de ses propres déterminations. Ce n'est qu'en adoptant un point de vue extérieur (*external standpoint*) que nous pouvons ainsi définir objectivement nos propres possibles. Or nous finissons toujours perdants de cette recherche des causes puisqu'il y aura toujours des déterminations de nos déterminations de nos déterminations, etc. : « le processus qui débute comme un moyen d'accroître la liberté semble conduire à sa destruction » (Nagel, 1993, p. 144). L'être humain ne peut être libre qu'à condition qu'il adopte sur lui-même un point de vue interne (*internal standpoint*) en se considérant comme l'origine d'actions possibles (et non probables).

En somme nous voyons ici, à partir de cet argument de Nagel, que la déconstruction de la seconde illusion – le moi comme agrégateur de données – nous conduit vers la première – le moi comme agrégat de données. La perspective d'une synthèse de soi par agrégation dynamique de ses données conduit à questionner ce moi qui agrège et qui ne se réduit pas à ses données. Or cette intentionnalité du moi peut elle-même être mise en données de sorte que le moi est de nouveau considéré comme un agrégat de données...

2.3. L'identité par les savoirs

Il semblerait donc que les données personnelles ne constituent pas des matériaux identitaires. Ce point renvoie à la distinction établie plus haut entre données et savoirs : il existe entre les données – brutes, informes et dénuées de sens – et les savoirs – élaborés, informés et pourvus de sens – une différence de nature et non de degré. Les savoirs ont fait l'objet d'une appropriation identitaire ; ils font

partie intégrante de la personne qui les possède et les partage au sein d'une communauté de savoirs.

Un exemple permettra d'apprécier cette différence fondamentale. Il nous est inspiré par le travail du collectif Dingdingdong autour de la maladie de Huntington, maladie neuro-évolutive à transmission dominante pour laquelle il existe depuis 1993 un test pré-symptomatique (Solhdju, 2015). Même dans ce cas où *a priori* les données livrées par le test-diagnostic pourraient sembler décisives, elles n'ont pourtant aucun sens en elles-mêmes. C'est toujours l'interprétation de ces données par des savoirs formés antérieurement qui permet de construire du sens. En l'occurrence, il est possible d'interpréter l'annonce du diagnostic comme une sentence, la maladie comme une malédiction et son évolution comme une dégradation catastrophique de l'existence. Mais il est également possible de considérer la maladie comme une « métamorphose » identitaire. C'est à ce titre que le collectif Dingdingdong défend la formation d'une communauté de savoirs constitués par les personnes concernées par cette maladie, qu'elles soient diagnostiquées ou non, porteuses ou non. C'est précisément par cette forme-ci d'*empowerment* que des savoirs collectivement formés peuvent maintenir non seulement la liberté d'exister mais aussi le sentiment identitaire des personnes concernées.

3. LA RECHERCHE D'UNE SOLIDARITÉ PAR AGRÉGATION DE DONNÉES

Nous souhaitons, dans ce dernier moment politique, considérer la perspective d'une synthèse du lien social par agrégation de données ou, dit autrement, de ce que nous nommerons provisoirement une « solidarité par les données ». Le principe est le suivant : par la mise en partage de leurs données multiples au sein de bases de données ouvertes, les membres d'une même société pourraient contribuer à une meilleure connaissance des déterminations individuelles, ce qui leur permettrait de tracer d'autant mieux leurs trajectoires de vie conformément à ce qu'ils désirent ou redoutent. Nous voyons bien comment cette perspective – dont nous montrerons les illusions – se bâtit sur la précédente : la « solidarité par les données » permet d'enrichir le profilage qui à son tour la renforce. Bienfaits collectifs et bienfaits individuels vont de pair : chacun peut appréhender précisément pour lui-même les risques d'une action entreprise dans un contexte donné ; mais il ne le peut qu'à condition d'avoir souscrit à un partage de ses propres données et à leur usage au sein d'algorithmes. De cette nouvelle forme de mutualisation, François Ewald vante les avantages dans un article récent (Ewald, 2013) : « Les données concernant chacun s'enrichissent de celle des autres et ceci en principe à l'infini dans la mesure où le partage des données engendre un surcroît de bien-

être. On a longtemps pensé que les nouvelles technologies de l'information allaient renforcer les tendances à la démutualisation. (...) Le monde des données peut servir de support à une nouvelle théorie de la mutualité, assez proche de la théorie des *commons*, c'est-à-dire de ces biens qui se valorisent d'être partagés (Wikipedia). Dans la mesure où les données sont des biens dont la valeur s'accroît avec leur partage, on peut penser que la mutualisation des données ne doit pas être restreinte par le régulateur ».

3.1. De la solidarité par les risques à la solidarité par les données ?

Dès lors un nouveau contrat se ferait jour entre l'individu et la collectivité qui implique, au regard d'un système de protection sociale bâtie sur la mutualisation des risques, plusieurs déplacements. Le premier concerne l'objet même du partage : il s'agit moins de protéger l'individu contre des épreuves vécues – la maladie, l'incapacité, le chômage, la retraite – en mettant en partage des ressources communes que de le protéger contre l'incertitude, l'errance ou l'ignorance, autant que les données disponibles le permettent.

Le deuxième déplacement concerne le risque, requalifié comme individuel. En premier lieu, il n'existe plus, tendanciellement au moins, de risques collectifs mais seulement des risques correspondant à un profil donné. En second lieu, puisque l'individu peut accéder à son profil, il devient responsable de ce qu'il en connaît et des conséquences qu'il en tire. Si l'ignorance et l'incertitude demeurent des choix possibles, l'individu en est désormais comptable. La délégation pure et simple de nos responsabilités à l'assurance sociale ou à l'Etat qui, éventuellement, pourraient parer aux coups du sort n'est plus envisageable. Dit autrement, la « solidarité par les données » implique de « déchirer le voile d'ignorance ». Par cette formule devenue célèbre, Pierre Rosanvallon désignait déjà en 1995 le processus par lequel la disponibilité des données individuelles pourrait rendre obsolète et contraire aux intérêts individuels le maintien d'un voile d'ignorance sur ces mêmes données (Rosanvallon, 1995). Ce gisement de données, notamment parce qu'il est doté d'une forte valeur économique, transforme la carte des intérêts. Les individus peuvent, potentiellement à profit, dessiner leurs parcours de vie ; quant aux assurances publiques ou privées, aux employeurs, aux banques, ils ont spontanément intérêt à connaître ces données pour ajuster leurs comportements.

En dernier lieu, le système de santé ne comprendrait plus en lui-même les activités du soin : en effet, si le soin implique la prise en compte d'une « vulnérabilité » par un tiers de confiance, alors force est d'admettre qu'un système de santé entièrement régi par l'utilisation des données de santé ne serait plus un système de soins à proprement parler, autrement dit un ensemble de dispositifs organisés pour soigner les individus, mais un système de données, capable de les recueillir,

de les stocker et de les croiser. Les activités de soin sont alors exfiltrées du « système de santé » et annexés à un domaine d'activité autonome.

3.2. La synthèse du lien social par agrégation ?

Nous ne souhaitons pas dire ici que la « solidarité par les données » est insuffisante et qu'elle ne peut se substituer à la solidarité par les risques ; nous contestons plus radicalement que la mise en partage de données puisse créer par elle-même le moindre lien social, la moindre *solidarité*. En somme c'est l'idée même d'une « solidarité par les données » que nous mettons en cause, en entendant la solidarité, à la suite de Serge Paugam commentant la pensée de Durkheim, comme un processus visant l' « intégration sociale », « l'intégration des individus à la société et l'intégration *de* la société » (Paugam, 2015), un processus par lequel quelque chose comme une société existe, se tient par elle-même et poursuit des fins collectives.

Les données ne sont en effet que les résidus individualisés et neutres issus d'un effort de décontextualisation (Rouvroy, 2014) et par là même de *séparation*. De même qu'à l'échelle du profil, l'agrégation s'accompagne de la fragmentation, elle s'accompagne, à l'échelle du collectif, de la séparation. Pour que les données puissent s'agréger entre elles, il a fallu considérer les personnes indépendamment de leur contexte de vie ; et il a fallu, pour ce faire, couper les relations réelles, faites d'affects, qui relient des personnes entre elles. Or ce sont précisément ces relations d'affects qui constituent le terreau vivant sur lequel se négocie, dans la conflictualité des pouvoirs, des valeurs et des goûts, la conscience d'appartenir à une même société.

Là encore prenons un exemple. Dans le monde des données, un phénomène comme l'obésité se traduit par la mesure d'un poids ; cette donnée pourra être corrélée à des données comportementales – alimentation, sédentarité, etc. – ; à des données socio-économiques – revenu, emploi, conditions matérielles d'existence ; à des données géographiques – espaces verts, disponibilités de lieux d'exercice physique, etc. ; à des données biologiques notamment génomiques. Cette compréhension ignorera cependant les processus d' « incorporation de la structure sociale » (Fassin, 2006), autrement dit l'ensemble des forces sociales affectant inégalement les corps, et notamment l'ensemble des représentations sociales rendant acceptable l'obésité des uns et intolérable celle des autres.

Mise en données, l'obésité demeure certes un phénomène « social », mais ce « social » est la recomposition d'un ensemble de faits neutres et vides d'affects. Sur cette base, la prise en compte politique de l'obésité comme « problème de santé publique » ne peut qu'obéir à une visée médico-économique – diminuer le « fardeau » de l'obésité – et ne peut que procéder de l'ingénierie sociale – « res-

ponsabiliser » les usagers, construire des équipements sportifs, mettre en place des incitations économiques, etc. – et non d'un mouvement de solidarité interrogeant l' « ordre économique et culturel de la société » (Genolini *et al.*, 2010). Plus encore, le fait même que l' « obésité » pose problème ne peut être remis en question alors même qu'il révèle de forts implicites culturels et notamment, selon Durif-Bruckert, « le rêve d'un corps vidé de sa matière qui pourrait se dégager de toutes traces liées aux opérations organiques : traces du besoin, de la dépendance, de l'altérité ».

Il ne s'agit pas ici de mettre en cause le partage des données lui-même ni les algorithmes dont la puissance peut être mise au service de finalités diverses, y compris critiques, comme l'ont bien montré les auteurs de l'ouvrage *Stactativisme* (Bruno *et al.*, 2014). Toutefois, même dans ces cas, le partage des données constitue moins un moment de solidarité, d' « intégration », qu'un moment utilitaire par lequel l'individu ou un collectif (déjà solidarisé) décuplent leur capacité à prédire l'avenir et à le déterminer au moyen d'algorithmes statistiques. Conférer au partage des données la fonction d'un processus de solidarité et, plus encore, d'un nouveau modèle de solidarité revient à embrasser une certaine lecture de ce qu'est la société. Celle-ci devient un ensemble d'individus que seule relie entre eux l'appartenance à une même population de données, et qui ne peuvent être comparés qu'au moyen de similitudes et de différences abstraites. Du reste, si cette pseudo-solidarité finit par se substituer à la solidarité fondée sur la mutualisation des risques, la société ne devra à l'individu que la mise à disposition d'un système de partage des données. Les inégalités deviendront de simples *différences* qu'il appartiendra à l'individu de modifier s'il le souhaite.

CONCLUSION : LE MÉPRIS DE L'EXISTENCE NON-NUMÉRIQUE

Nous n'avons pas souhaité critiquer directement le phénomène *big data* mais davantage avons-nous souhaité porter l'attention sur l'illusion ontologique qui accompagne ce phénomène, illusion selon laquelle nos connaissances, nos identités, nos liens pourraient se constituer à partir d'opérations d'agrégations. Les données se présenteraient comme les outils d'une nouvelle intelligence du monde, de soi et des autres ; elles permettraient de donner un contenu et des frontières à des essences que nous ne parvenons pas à saisir. En somme les données seraient les nouveaux donneurs de sens. Une telle illusion, dont on peut bien concevoir la force de séduction, nous semble dangereuse moins par ses effets que par le partage qu'elle établit entre ce qui peut être mis en données et ce qui ne peut l'être. À moins de confondre singularité et particularité, sentiments et émotions, les éléments contextuels de l'existence ne peuvent être numérisés. Ils res-

teront toujours en marge des bases de données. Une vision numérique du monde est donc empreinte d'une forme de mépris qui pourra se traduire concrètement par la marginalisation de certaines disciplines, la réduction des identités à des profils ou encore l'invisibilisation des conflits sociaux.

RÉFÉRENCES BIBLIOGRAPHIQUES

Audard C. (2016), « Les frontières de la justice : citoyenneté, capacité d'anticipation et handicap », *Revue française d'éthique appliquée*, numéro 2.

Anderson C. (2008), « The end of theory : the data deluge makes the scientific method obsolete », *Wired* [<https://www.wired.com/2008/06/pb-theory>].

Babinet G. (2015), *Big data, penser l'homme et le monde autrement*, Paris, Le Passeur.

Bachelard G. (1965), *L'activité rationaliste de la physique contemporaine*, Paris, PUF.

Benasayag M. (2016), *Cerveau augmenté, homme diminué*, Paris, La Découverte.

Bollier D. (2010), *The promise and peril of Big Data*, Report of the The Aspen Institute.

Boyd D. & Crawford K. (2012), « Critical Questions for Big Data », *Information, Communication & Society*, 15-5 : 662-679.

Bruno I., Didier E., Prévieux J. (Éds.) (2014), *Statactivisme. Comment lutter avec les nombres*, Paris, Zones.

Coutellec L. (2015), « Du big data au big ficta : un itinéraire scientifique, technique et social à partager », *Les Cahiers de l'Espace éthique*, 2 : 64-67.

Frydman B. & Van Waeyenberge A. (Eds) (2013), *Gouverner par les standards et les indicateurs. De Hume aux rankings*, Bruxelles, Bruylant.

Dardot P. & Laval C. (2009), *La nouvelle raison du monde. Essai sur la société néolibérale*, Paris, La Découverte.

Darrason M. (2014), « Y a-t-il une théorie génétique de la maladie ? », thèse soutenue le 2 juillet 2014 à l'Université Paris1-Panthéon-Sorbonne.

deCODEme (2009), "deCODEme Complete Scan," En ligne [<https://www.decode.com/complete-genetic-scan>], dernier accès : août 2009.

Desrosières A. (2008), *Gouverner par les nombres*, Paris, Mines Paristech.

Desrosières A. (2014), *Prouver et gouverner. Une analyse politique des statistiques publiques*, Paris, La Découverte.

Durif-Bruckert C. (2017), « On devient ce que l'on mange : les enjeux identitaires de l'incorporation », *Revue française d'éthique appliquée*, 4 : 25-36.

Fassin D. (2006), « Le sens de la santé. Anthropologie des politiques de la vie », dans F. Saillant et S. Genest (dir.), *Anthropologie médicale. Ancrages locaux, défis globaux*, Québec, Les Presses de l'Université Laval, pp. 383-399.

Fox Keller E. (2003), *Le siècle du gène*, traduction par Stéphane Schmitt, Paris, Gallimard.

Génolini J, Clément J (2010), « Lutter contre la sédentarité : L'incorporation d'une nouvelle morale de l'effort », *Sciences sociales et sport*, 1-3 : 133-156.

Grimaldi N. (2001), *L'Homme disloqué*, Paris, Presses Universitaires de France.

Hey T. et al. (2009), *The fourth paradigm. Data-intensive scientific discovery*, Microsoft Research.

Juengst E., Flatt M., Setters M. (2012), « Personalized Genomic Medicine and the Rhetoric of Empowerment » *Hastings Center Report*, Sept-Oct. 2012. En ligne [<http://onlinelibrary.wiley.com/doi/10.1002/hast.614/abstract>].

Lamy J. (2007), « Penser les rapports entre sciences et politique : enjeux historiographiques récents », *Cahiers d'histoire*, 102 : 9-32.

Mill J.-S. (1990), *De la liberté*, traduction par Lenglet L., Paris, Folio.

Nagel T. (1993), *Le point de vue de nulle part*, traduction par Sonia Kronlund, Paris, Editions de l'éclat.

Paugam S. (Éd.) (2011), *Repenser la solidarité. L'apport des sciences sociales*, Paris, Presses universitaires de France.

Rawls J. (1987), *Théorie de la justice*, traduction par Audard C., Paris, Seuil.

Rosa H. (2012), *Accélération et aliénation. Vers une théorie critique de la modernité tardive*, Paris, La Découverte.

Rosanvallon P. (1995), *La nouvelle question sociale*, Paris, Seuil.

Rouvroy A. (2010), « Le nouveau pouvoir statistique », *Multitudes*, n°40.

Rouvroy A. (2014), « Des données sans personne : le fétichisme de la donnée à caractère personnel à l'épreuve de l'idéologie des Big Data », *Etude annuelle du Conseil d'Etat: le numérique et les droits fondamentaux* (p. 407-422), Paris, La Documentation française.

Rouvroy A. & Stiegler B. (2015), « Le régime de vérité numérique. De la gouvernabilité algorithmique à un nouvel Etat de droit », *Socio*, n°4.

Solhdju C. (2014), *L'épreuve du savoir*, Paris, Les presses du réel.

Simondon G. (2012), *Du mode d'existence des objets techniques*, Paris, Aubier [1958].

Stiegler B. (2015), *Dans la disruption. Comment ne pas devenir fou ?*, Paris, Les liens qui libèrent.

Supiot A. (2015), *La gouvernance par les nombres*, Paris, Fayard.