



UNIVERSITAT DE  
BARCELONA



Observatori de  
Bioètica i Dret  
Universitat de Barcelona



FLACSO  
ARGENTINA

Revista de Bioética y Derecho

Perspectivas Bioéticas

www.bioeticayderecho.ub.edu - ISSN 1886-5887

## DOSSIER BIG DATA

**Bioethics in the big data era: health care and beyond**

**La bioética en el época del 'big data': la salud y más allá**

**La bioètica en l'època del 'big data': la salut i més enllà**

**SARAH CHAN \***

## OBSERVATORI DE BIOÈTICA I DRET DE LA UNIVERSITAT DE BARCELONA

La Revista de Bioética y Derecho se creó en 2004 a iniciativa del Observatorio de Bioética y Derecho (OBD), con el soporte del Máster en Bioética y Derecho de la Universidad de Barcelona: [www.bioeticayderecho.ub.edu/master](http://www.bioeticayderecho.ub.edu/master). En 2016 la revista Perspectivas Bioéticas del Programa de Bioética de la Facultad Latinoamericana de Ciencias Sociales (FLACSO) se ha incorporado a la Revista de Bioética y Derecho.

Esta es una revista electrónica de acceso abierto, lo que significa que todo el contenido es de libre acceso sin coste alguno para el usuario o su institución. Los usuarios pueden leer, descargar, copiar, distribuir, imprimir o enlazar los textos completos de los artículos en esta revista sin pedir permiso previo del editor o del autor, siempre que no medie lucro en dichas operaciones y siempre que se citen las fuentes. Esto está de acuerdo con la definición BOAI de acceso abierto.

\* Sarah Chan. Usher Institute for Population Health Sciences and Informatics, The University of Edinburgh, United Kingdom. E-mail: [sarah.chan@ed.ac.uk](mailto:sarah.chan@ed.ac.uk).

## Abstract

'Big data' and data-intensive research approaches are rapidly gaining momentum in health and biomedical research, with potential to transform health at all levels from personal to public. The use of 'big data' for health research, however, raises a number of ethical challenges. In this paper I discuss ethical aspects of the advent of big data in health. I argue that although public discourse has focused on immediate concerns relating to use of individuals' information, 'big health data' requires us to explore alternative conceptual approaches to research ethics, including the 'social contract' model. Further, we need to think beyond health research uses of data to the social consequences of big data epistemology and practice, and the moral implications of 'datafying' the human.

**Keywords:** bioethics; big data; population health; data science; research ethics; genomics; ethics of algorithms; social media.

## Resumen

La ciencia de 'big data' (o datos masivos) lleva mucho potencial para la investigación biomédica, y promete una transformación en la salud y la asistencia médica. Al mismo tiempo, el uso de datos de salud en investigación presenta varios retos éticos. En este artículo, exploraré aspectos éticos de la llegada del 'big data' al ámbito de la salud. Aunque el discurso público y regulatorio se ha focalizado mucho en el uso de datos del individuo, lidiar con los nuevos desafíos de datos masivos requiere considerar enfoques alternativos a la ética de la investigación, tal como el modelo del "contrato social". Hay que pensar más allá del uso de datos para investigaciones en salud y contemplar las consecuencias sociales de la epistemología y la práctica de 'big data' y las implicancias morales de la 'datificación' del humano.

**Palabras clave:** bioética; big data; datos masivos; salud poblacional; datos de salud; ética de la investigación; genómica; ética de algoritmos; medios sociales.

## Resum

La ciència del 'big data' (o dades massives) comporta un enorme potencial per a la recerca biomèdica, i promet ocasionar una gran transformació en l'àmbit de la salut i l'assistència mèdica. Al mateix temps, l'ús de dades de salut en recerca presenta diversos reptes ètics. En aquest article, analitzaré els aspectes ètics de l'arribada del 'big data' a l'àmbit de la salut. Encara que el discurs públic i regulador s'ha focalitzat principalment en l'ús de les dades personals, bregar amb els nous desafiaments que comporten la irrupció de les dades massives requereix enfocaments alternatius a l'ètica de la recerca, com ara el model del "contracte social". A més, cal pensar més enllà de l'ús de dades per a recerques en salut i tenir en compte les conseqüències socials de l'epistemologia i la pràctica del 'big data' i les implicacions morals de la 'datificació' d'allò que és humà.

**Paraules clau:** bioètica; big data; dades massives; salut poblacional; dades de salut; ètica de la recerca; genòmica; ètica d'algorismes; mitjans socials.

## 1. Introduction

Big data is changing the way we live. Every moment, we are generating digital data that can be collected, stored and used in myriad ways. For example when we engage in such everyday activities as searching the Internet, posting on social media, using location services on our smartphones or online shopping, we create virtual traces of ourselves that may persist, acquire permanence and have effects far beyond the action that itself generated the data. Equally, in what might be considered more sensitive spheres of personal life, such as when we go to the doctor or access social services, electronic records are generated that represent our interactions with the system. At potentially every moment, our existence in the world leaves trails of data footprints. What is done with this data – and by whom – has the capacity to transform our world.

One of the areas of data-intensive research that is currently seen as most promising is the use of big and smart health data. Increases in computing power, together with our growing ability to measure different aspects of human biology, allow for the collection of an ever-expanding quantity of highly varied data.

Aggregating and analyzing this data has the potential to produce new approaches to disease, diagnosis and treatment, public health, and medical research and innovation.

Big health data thus promises a revolution in health at all levels from individual care to public health<sup>1</sup>. At the same time, personal health information is regarded as potentially highly sensitive, provoking concerns about its use<sup>2</sup>. Added to this, the possible consequences of both the results of big data research and how they are used, as well as the conceptual and relational transformations entailed by this shift in our way of seeing the world, require ethical attention. The aim of this paper is therefore to explore the concept of big (health) data and the challenges it presents for bioethics, and to seek to identify some of the key issues and ideas we will need to address to meet those challenges.

---

<sup>1</sup> ACADEMY OF MEDICAL SCIENCES: Personal Data for Public Good: Using Health Information in Medical Research., Academy of Medical Sciences, London 2006; GROVES, PETER et al.: The 'Big Data' Revolution in Healthcare: Accelerating Value and Innovation, Centre for US Health System Reform Business Technology Office 2013; PRECISION MEDICINE INITIATIVE (PMI) WORKING GROUP REPORT TO THE ADVISORY COMMITTEE TO THE DIRECTOR, NIH.: The Precision Medicine Initiative Cohort Program – Building a Research Foundation for 21st Century Medicine. September 17, 2015.

<sup>2</sup> NUFFIELD COUNCIL ON BIOETHICS: The Collection, Linking and Use of Data in Biomedical Research and Health Care: Ethical Issues 2015; ACADEMY OF MEDICAL SCIENCES: Realising the Potential of Stratified Medicine 2013; LLÀCER, M R, CASADO, M, and BUISAN, L: Document on Bioethics and Big Data: Exploitation and Commercialisation of User Data in Public Health Care, Observatori de Bioètica i Dret, Barcelona 2015.

## 2. Current concerns

### 2.1 Big data for health: examples

What do we mean by referring to 'big (health) data'? Considering some examples perhaps gives us a flavour of the kinds of research under discussion:

Large-scale population genomic studies are one of the most obvious forms of current big data health research. Population health science has long involved the assembly and analysis of relatively large quantitative datasets. Genomics, while a more recent science, has been one of the primary drivers of the scaling-up of data processes in biological research, with the turn to 'big biology' first exemplified by the Human Genome Project. Genetic (or molecular) epidemiology combines these two approaches to analyse health records against genomic data, allowing the identification of genetic correlates of health and disease across a wide population.

The level of detail available for such studies is rapidly increasing from genome-wide association studies using markers such as SNPs, to whole-genome sequencing initiatives. The 100,000 Genomes Project<sup>3</sup>, for example, aims to collect whole-genome sequence data from patients with rare disease and their families, and cancer patients, with the dual objectives of improving clinical care via 'genomic medicine' and producing new knowledge about the molecular basis of disease. Population genomic approaches can also be combined with biobanking, linking physical bioresources to the genetic and health information to enable the study of a wide range of complex biological attributes.

Another enabling factor in big health data is the growing use of electronic health records (EHRs), transforming patient health information into potentially usable datasets. The Scottish health system (NHS Scotland), for example, maintains comprehensive records with a unique identifier for each patient, which together with administrative and social data collected by National Services Scotland, creates the possibility for wide-ranging studies linking health and social data across the entire population. Research of this sort has produced important findings in individual and public health, such as in relation to the health consequences of obesity in pregnancy, the treatment of acute pancreatitis and the public health impact of anti-smoking legislation.

---

<sup>3</sup> <https://www.genomicsengland.co.uk/the-100000-genomes-project/>.

These and other examples illustrate how depth as well as breadth of data, together with the analytical tools and processing capacity to handle the increasingly complex data matrices produced, contribute to the transformative potential of big data for population health.

Big data also promises transformative effects on individual clinical care, via the advent of personalized and precision (or stratified) medicine approaches. Lee Hood, a pioneer and proponent of “systems medicine”, as he calls it, envisages “every consumer of health care surrounded by a virtual cloud of billions of data points”<sup>4</sup>. This data cloud will include medical information from health records, genetic, genomic, proteomic and other molecular data, and data relating to social and environmental factors. As well as facilitating research, the data can be used to improve personalized predictive and preventive care, such as through biomarker monitoring to detect pre-disease progression and allow early intervention, and targeted (“personalised”) approaches to treatment. Molecular profiling of cancers to determine the likely effectiveness of various therapies, for example, is increasingly being incorporated into treatment protocols.

Big data may also have public health applications in epidemiology and response to disease<sup>5</sup>. Importantly, this is not limited to information directly associated with the health care or biomedical context: during the Ebola epidemic, for example, mobile phone data helped to track the movement of persons in and around affected areas, and hence predict and respond to the spread of disease<sup>6</sup>.

Using social data in this way is potentially very powerful, but can also be fallible, as shown by the following much-reported example. In 2008 Google launched the disease monitoring tool Google Flu Trends (GFT), based on tracking search query terms whose frequency had been shown to correlate with rates of influenza-like illness reported by the US Centers for Disease Control and Prevention (CDC) using traditional means of surveillance. It seems reasonable that people suffering from influenza-like symptoms might turn to internet search engines to look for diagnostic advice and remedies, and that therefore analyzing patterns of search terms could provide a form of disease surveillance that would help to track seasonal flu outbreaks. Indeed, for

---

<sup>4</sup> HOOD, LEROY and FLORES, MAURICIO: "A Personal View on Systems Medicine and the Emergence of Proactive P4 Medicine: Predictive, Preventive, Personalized and Participatory," *New Biotechnology* 29, no. 6, 2012, 613-24.

<sup>5</sup> DOWELL, SCOTT F, BLAZES, DAVID, and DESMOND-HELLMANN, SUSAN: "Four Steps to Precision Public Health," *Nature* 540, 2016, 189-91.

<sup>6</sup> WALL, MATTHEW: "Ebola: Can Big Data Analytics Help Contain Its Spread?," *BBC News*, 15 October 2014. <http://www.bbc.co.uk/news/business-29617831>, accessed 1 June 2017; TALBOT, DAVID: "Cell-Phone Data Might Help Predict Ebola's Spread," *MIT's Technology Review*, August 22 2014.

<https://www.technologyreview.com/s/530296/cell-phone-data-might-help-predict-ebolas-spread/>, accessed 1 June 2017.

a number of years, GFT was able to deliver estimates that very closely matched the CDC's data. During the 2012-13 season, however, the results delivered by GFT vastly over-predicted the reported incidence of disease, estimating an infection rate almost double that actually reported<sup>7</sup>.

This discrepancy should not be read as indicating that social media is an inherently unsuitable source for big data health research, more that the algorithm was not optimized to deal correctly with the data received; that is, it had not been 'trained' so as to account adequately for the complex multitude of factors influencing the population's Google search patterns in this situation. Refining the algorithm iteratively against data collected in diverse circumstances will increase its power to make accurate predictions. Nevertheless, this example indicates the need to remain cautious and critical about how much we rely on big data and algorithmic predictions to tell us about the state of the world.

## 2.2 What's 'big' about 'big data'?

As can be seen from the examples given above, the range of questions, topics and approaches that fall under the definition of 'big data' or 'data-intensive research' is at once virtually limitless and highly varied. Indeed, it may be that there is no such thing as a singular, homogeneous 'big health data science'; differences between sub-fields, for example with respect to methods of collection, processing and analysis, may mean that modes of research grouped together under this heading turn out to be quite different in philosophy and practice. Nevertheless, treatments of the concept of 'big (health) data' have tended to pick out certain key characteristics that might be said to be distinctive, and also give rise to particular ethical features.

Early discussions described big data in terms of the "3Vs": volume, variety and velocity<sup>8</sup>. 'Big data' is big in the sense of being more than can be manually analysed, requiring the development of novel computational approaches to make sense of the analysis. It is big in the sense of variety, combining data from multiple sources: those relevant to health include not only patients' EHRs but social and administrative data, together with data from research and development, consumer information and more. Moreover, this variety also includes new data

---

<sup>7</sup> BUTLER, D.: "When Google Got Flu Wrong," *Nature* 494, no. 7436, 2013, 155-6.

<sup>8</sup> FULLER, MICHAEL: "Big Data: New Science, New Challenges, New Dialogical Opportunities," *Zygon* 50, no. 3, 2015, 569-82.

forms, via what Cukier and Mayer Schonberg call 'datafication': "the ability to render into data many aspects of the world that have never been quantified before"<sup>9</sup>.

Variety also refers to multiple possible forms of input: these include personal devices and wearable technology, eHealth and social media, in addition to traditional ways of collecting health information. In this sense, big data is inextricably linked to the digital revolution, Web N.0 and the personalised cybersphere; in other words, who we are in the digital world. Finally, increases in processing power and our ability to measure all sorts of data in greater and greater detail, mean that the pace at which big data is collected is rapidly accelerating.

Alongside these features, big data involves claims about altered epistemologies and novel practices. Leonelli<sup>10</sup> considers three features that have been identified as key shifts that characterize big data<sup>11</sup>:

1. "comprehensiveness", the idea of assembling all the data about a phenomenon, or as much of it as possible, to enable our analysis ("n = all"<sup>12</sup>);
2. "messiness", the idea that we can sacrifice more controlled, accurate and targeted data collection in return for much greater volume that can deliver a better overall picture rather than painstaking detail; and
3. a shift in the kind of knowledge that is seen as important, "from causation to correlation"<sup>13</sup>, whereby correlation "comes to be appreciated as not only a more informative and plausible form of knowledge than the more definite but also a more elusive, causal explanation"<sup>14</sup>.

In relation to biology, Leonelli contends that big data may be less of an epistemic revolution than is claimed, but nonetheless involves novel orientations of practice, notably practices that create value (or permit value to be created) in data; and for the handling and analysis of data. This new approach to data as having value that can be realized through varied modes of analysis is likely also to translate into other areas where 'big data' epistemology is having more of an impact,

---

<sup>9</sup> CUKIER, KENNETH and MAYER-SCHOENBERGER, VIKTOR: "The Rise of Big Data: How It's Changing the Way We Think About the World," *Foreign Affairs* 92, no. 3, 2013, 28-40.

<sup>10</sup> LEONELLI, SABINE: "What Difference Does Quantity Make? On the Epistemology of Big Data in Biology," *Big Data & Society* 1, no. 1, 2014, 1-11.

<sup>11</sup> MAYER-SCHÖNBERGER, VIKTOR and CUKIER, KENNETH: *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray, 2013; CUKIER, KENNETH and MAYER-SCHOENBERGER, VIKTOR: "The Rise of Big Data: How It's Changing the Way We Think About the World".

<sup>12</sup> ———: "The Rise of Big Data: How It's Changing the Way We Think About the World".

<sup>13</sup> Ibid.

<sup>14</sup> LEONELLI, SABINE: "What Difference Does Quantity Make? On the Epistemology of Big Data in Biology", 3.

such as social and economic activity. Big health data, lying at the intersection of biology and these wider spheres of data use, is no exception. The Nuffield Council report on data in health care and biomedicine recognizes data as “a valuable resource that may be reused indefinitely in other contexts, linked, combined or analysed together with data from different sources...”<sup>15</sup>.

Taken all together, these features of big health data are what lead to particular ethical concerns.

### 2.3 Current concerns in big data research

Big data health research, in particular the use of EHRs and other patient data, provokes a range of immediate conventional concerns about the use of individuals’ personal information: to whom does it belong, who can access it, how can it be used? Such concerns manifest in terms of the interrelated concepts of privacy, confidentiality and consent<sup>16</sup>, which focus on individuals’ interests in their own personal information. Additional questions arise over what constitutes personal information, and who has or should have control of this for different purposes: disclosure, use or even commercial gain. De-identification may make data less personal, but there is a trade-off in that the utility of data depends on being able to link it to other information about that individual – genomic, demographic and health information, for example – which in turn increases the possibility of re-identification. When personal data is held there are also worries over data protection and what happens if a breach of data security occurs.

These concerns have always been present in relation to personal information, but the characteristics of big health data will potentially exacerbate them. In relation to volume and variety, the ability to collect and connect a much wider range of data and use it for health-related purposes may create new ways in which privacy can be infringed, and new harms to which people may thus be exposed.

Consider for example the implications for privacy of monitoring behaviour via eHealth technologies. Electronic pedometers, or step counters, have become increasingly popular as a way of allowing people to monitor their physical activity and, presumably, encourage them to be more active. Sales of Fitbit, a wearable device that monitors activity levels and other biometrics such as heart rate and sleep, more than doubled each year from 2010 to 2015<sup>17</sup>; Apple iPhones now come

---

<sup>15</sup> NUFFIELD COUNCIL ON BIOETHICS: *The Collection, Linking and Use of Data in Biomedical Research and Health Care: Ethical Issues*, 15.

<sup>16</sup> For a discussion of these in relation to health data see *ibid*.

<sup>17</sup> <https://www.statista.com/statistics/472591/fitbit-devices-sold/>, accessed 30 May 2017



with a built-in pedometer app. Other digital technologies might track eating behaviour, for example by tracking food purchases, refrigerator contents and people's movements<sup>18</sup>.

In a world where health, fitness and fatness are increasingly moralized, information about one's daily level of physical activity or how many times one has opened the fridge for a snack might well be considered sensitive and personal. Indeed, some apps have traded on exactly that, attempting to discourage "undesirable" behaviour by posting reports on users' late-night snacking to social media<sup>19</sup>. Friends are encouraged to join in by commenting and shaming the midnight snacker for their lapse.

The consequences of such monitoring, however, can go further than merely being subjected to unwanted judgmental attitudes from friends or society at large. Could this data be used in treatment decisions, to ensure compliance with medically recommended behavioural regimens? Access to surgical procedures is already in some cases conditional on weight loss or smoking cessation; we might imagine overweight patients being refused treatment because doctors deem on the basis of their fridge visits and exercise habits that they have not made a sufficient effort to shape up.

We might regard this sort of 'dataveillance' for health as an unacceptable level of state intrusion into the private domain. Others might argue that holding individuals to account for their private actions in this way is justified because of the public health benefits, or because it will result in fairer distribution of health resources. Big health data prompts us to develop new accounts of 'health privacy', to question to what extent we should blame or absolve people of health responsibility for actions that were previously private and can now be revealed, and how we should use data to do so.

In addition, the newly-recognised value associated with big data gives rise to an information marketplace, requiring attention to how value is created in data, how that value is realized and to whom it flows. These issues are also linked with questions of access to and control over data: who owns a dataset? Who has the right to grant or deny access, or to profit from its use?

When it comes to big data, these concerns go beyond the personal, individual level. While data now has scientific, economic and social value, individuals' data is valuable primarily in virtue of its contribution to the collective. One single person's health data tells us nothing in terms of generalizable inferences; it is in the context of the whole, and how that whole is curated to allow meaningful analysis, that it takes on value. This being the case, how should we regard our role in

---

<sup>18</sup> For example the app 'Carrot Hunger'.

<sup>19</sup> One such example is the app 'Virtual Fridge Lock'.

relation to our own data and its contribution to the collective? Are we shareholders of an economically valuable asset, or are we joint owners or perhaps stewards of a public good or common resource? What new relationships —among people, populations, health care providers, researchers and companies —are created by the use of big data, or further, by its commercialization?

Another set of current issues over big health data relate to interoperability and ethical governance, and the practices required to achieve this. This includes determining information standards for how data can be made usable and re-usable, for example in terms of formatting and metadata. While this seems like a scientific, rather than an ethical issue, how these standards are set will affect which data gets incorporated into big data research. Paradoxically, the standards required to make data 'Big' may make smaller the pool of data that might be counted as Big, at least in certain contexts. Particularly in scientific practice, where data is still purposively generated more than incidentally harvested, there may be expectations around data standards that have an exclusive effect.

Without keen attention to the factors that shape this, we may miss ethically-relevant consequences. For example, if standards imposed by the scientific community make it more difficult for certain groups to prepare their data in a compliant way, then data from those groups risks being excluded. This may affect the validity of research derived from selective datasets. We know that the disproportionate representation of "WEIRD" people in psychological and social research<sup>20</sup> limits the wider applicability of many findings, while lack of diversity amongst clinical trial participants has serious consequences for the generalizability of treatments to the population at large; similar caveats will apply to data science done on a skewed dataset. Constraining whose data is permitted to become part of 'big data' also has implications for scientific justice, in terms of barriers to participation in science and whose voices are represented and recognized in the scientific community.

In terms of big health data, the primary purpose of collecting patient information in the clinical context has so far been to improve care. Certain standards are required to make information usable for this purpose; optimizing its usability for research may require additional measures, as well as infrastructure and resources for set-up and administration of databases. A question for big health data ethics, then, is what if any changes may be required to achieve 'big datafication' of routine health information, and the effects of these on inclusivity of datasets as

---

<sup>20</sup> HENRICH, JOSEPH, HEINE, STEPHEN J, and NORENZAYAN, ARA: "Most People Are Not Weird," *Nature* 466, no. 7302, 2010, 29; ———: "The Weirdest People in the World?," *Behavioral and Brain Sciences* 33, no. 2-3, 2010, 61-83; discussion 83-135.

well as on health care practice. Additionally, an important consideration for future global health will be the development of health data science capacity in low and middle-income countries, to ensure that the benefits of big health data are available to these populations.

Finally, the epistemology of big data also presents a challenge for ethical standards and research integrity. By definition, it is not always easy or possible to see the ramifications of big data research from looking at the data with human eyes. How, then, can we ensure responsibility and integrity in the wider sense, when analysis & decision-making is delegated to machines?

### 3. Research ethics for the data era

#### 3.1 New approaches to research ethics

Many of the current concerns over health data research foreground the individual as the source and subject of data-driven research. Big health data, however, in going beyond the individual to the level of publics and populations, requires us to reconceptualise our roles with respect to "our" data, health care and research, and to develop new frameworks for the ethics of research using big health data to account for these altered relationships.

If we consider the development of bioethical thinking about human participant research over the past 70-odd years, it is evident that the concerns and problems that were foremost in the past have shaped research ethics towards an emphasis on protecting participants from the possible harms of research. This has led to a precautionary approach to governance and a focus on informed consent, at times to the exclusion of almost all else. Such an approach is understandable given that this framework developed largely in response to historical abuses of research participants in the biomedical context, but has left us with the idea that research is something inherently harmful or dangerous, from which participants must be protected; and that the ethics of research must therefore be different to the ethics of clinical care – that research and treatment should be considered separate. Both of these assumptions require re-evaluation, especially in the context of health data research. The principles of protecting participants and respecting autonomy are still important, but they are not, perhaps, the principles that are most at stake when it comes to big health data.

Big data demands a new ethical approach, to health care and health research – which are increasingly becoming part of the same process, or at least beginning to overlap. These changes have already begun to manifest across biomedicine, health care and innovation; 'big data' research, though, brings them into sharp focus. When every visit to a hospital or GP produces data

that feeds into research, the putative separation between research and treatment begins to seem untenable. Further, although there are possible harms that can occur as a result of big data research participation, they are of a different nature to those incurred in the course of clinical research that involves direct bodily intervention such as administering a new drug or procedure, and feature a different balance and distribution of risk against the potential benefits.

It is widely acknowledged that we need a new way of thinking about the ethics of research participation, in order to navigate the evolving terrain of health care, biomedicine and health innovation. From the patient perspective, in line with the shift from paternalism to autonomy in medical practice, this landscape is one in which patients are increasingly placed as active agents making (supposedly) autonomous choices about treatment and participation, rather than passive recipients of treatment in their best interests. Patients play a growing role in driving science through both political and consumer demand: the desire for new treatments propels political movements to enable access, such as 'right-to-try' legislation, as well as creating a market for health innovation. This new role is additionally facilitated by the digital age, via increased connectivity such as through social media, and increased access to information.

On the science side, new forms of health research challenge the appropriateness and practicability of a model focused on individual consent to specific studies: should research using stored biomaterials, for example, require tissue progenitors to consent as participants to each use? Big health data, of course, is a prime example of research that sits uneasily with the existing ethical paradigm.

### 3.2 Data science and the social contract

To deal with the challenges posed by new forms of research and new modes of participation, bioethicists have begun to turn to a "social contract" approach to health care and research<sup>21</sup>. Such an approach need not mean a radical overhaul of the norms and values underlying the previous model, more a reorientation that prompts us to re-examine some of the embedded assumptions about research and what makes it ethical. In alignment with the arguments some have raised that

---

<sup>21</sup> CHAN, SARAH, HARRIS, JOHN, and SULSTON, JOHN: "Science and the Social Contract: On the Purposes, Uses and Abuses of Science," in *Common Knowledge: The Challenge of Transdisciplinarity*, ed. Jerome Billotte, et al. (Lausanne: EPFL Press, 2010); MESLIN, E. M. and CHO, M. K.: "Research Ethics in the Era of Personalized Medicine: Updating Science's Contract with Society," *Public Health Genomics* 13, no. 6, 2010, 378-84; HORNE, ROB et al.: "A New Social Contract for Medical Innovation," *Lancet* 385, no. 9974, 2015, 1153-4; VAYENA, EFFY et al.: "Research Led by Participants: A New Social Contract for a New Kind of Research," *Journal of Medical Ethics* 42, no. 4, 2016, 216-9.

there may be moral reasons to participate in research<sup>22</sup>, it seeks to shift the presumption that research participation is necessarily harmful or significantly burdensome, and to recognize the important benefits that can flow from research and that research not-done represents an opportunity cost.

On the social contract model, science is characterized as a valuable social institution that conduces to public benefit, human welfare and the good functioning of society. It gives rise to rights (expectations) regarding the benefits of science and how they should be distributed, as well as responsibilities (obligations) in relation to science, including the obligation to support and contribute to research. Participation in research, especially where the burdens are low and the risks proportionate, should therefore be seen less as supererogatory and more part of civic duty.

The social contract approach may be particularly applicable to big health data, for a number of reasons<sup>23</sup>. The first is interdependence: precision medicine is an enterprise whose value depends on wide participation, in which the ability to diagnose and determine the best treatment for individuals depends on the contribution of others. Likewise, individuals may seek access to precision care to benefit their own health, but their participation in the research enterprise is necessary in order to sustain the system that delivers these benefits. What we do with 'our' data affects others; big health data research can provide collective social benefits as well as serving individuals' interests in their own health.

Data research also generally imposes relatively low burdens on participants, especially where the data in question would be collected in any case as part of health care, and promises high potential benefits in return.

Moreover, the scope of big health data should prompt us to think towards the level of global society and the need for wider cooperation in research. Given the applicability, perhaps even necessity, of big data approaches to manage global health challenges such as pandemic disease, an ethical approach that explicitly admits socio-political framings will be useful.

In support of a social contract model of research, such an approach is congruent with commonly held public perceptions of the function of the health system and of research. The

---

<sup>22</sup> CAPLAN, ARTHUR L: "Is There a Duty to Serve as a Subject in Biomedical Research?," *IRB: A Review of Human Subjects Research* 6, no. 5, 1984, 1-5; RHODES, ROSAMUND: "Rethinking Research Ethics," *American Journal of Bioethics* 5, no. 1, 2005, 7-28; HARRIS, JOHN: "Scientific Research Is a Moral Duty," *Journal of Medical Ethics* 31, no. 4, 2005, 242-8; CHAN, SARAH and HARRIS, JOHN: "Free Riders and Pious Sons--Why Science Research Remains Obligatory," *Bioethics* 23, no. 3, 2009, 161-71.

<sup>23</sup> DESMOND-HELLMANN, SUSAN: "Toward Precision Medicine: A New Social Contract?," *Science Translational Medicine* 4, no. 129, 2012, 129ed3.

catchphrases that 'data saves lives' and 'rights require responsibilities' are gaining currency in the health care setting, acknowledging the moral duty of beneficence and that the right to receive health care and the benefits of ongoing improvements in medical knowledge is part of a system that also requires our contribution if it is to continue producing these benefits. Anecdotally, clinical researchers often report that patients' expectations within the health care system as to what is done with their information are concordant with routine incorporation of data into research<sup>24</sup>. Studies of views and experiences of research participants also show that they place higher importance on trust in researchers and the good that the research will achieve, than on the details of the consent process and precise information about each specific study<sup>25</sup>. Belief in the public value of science and the idea of fairness are important factors that support the legitimacy of research<sup>26</sup>.

Nevertheless, the focus of public concern and regulatory discourse with respect to health data research remains largely on consent, security, and individual-level control of data<sup>27</sup>. What is needed is a framework that is better able to take account of how big health data brings us into different roles with respect to each other, society, science and the state, or perhaps more obviously juxtaposes the roles we already simultaneously occupy.

### 3.3 Changing roles in the data era

Big data, among other emerging forms of health research, alters the space in which individuals are positioned with respect to the health care system, research and innovation. Within this, the roles available to citizens with respect to science are also shifting: are we patients, participants – or consumers?

Comparing two recent examples illustrates the complex dynamics and relationships at play in the management of health data. In 2012, the English National Health Service (NHS England) announced the *care.data* initiative, a scheme that proposed to make patients' NHS primary care

---

<sup>24</sup> A question this raises, of course, is whether patients with chronic or serious illnesses may in general be more willing for their data to be used in research than people who are mostly 'healthy'.

<sup>25</sup> DIXON-WOODS, MARY et al.: "Beyond "Misunderstanding": Written Information and Decisions About Taking Part in a Genetic Epidemiology Study," *Social Science and Medicine* 65, no. 11, 2007, 2212-22.

<sup>26</sup> DIXON-WOODS, MARY et al.: "Human Tissue and 'the Public': The Case of Childhood Cancer Tumour Banking," *BioSocieties* 3, no. 1, 2008, 57-80; DIXON-WOODS, MARY and TARRANT, C.: "Why Do People Cooperate with Medical Research? Findings from Three Studies," *Social Science and Medicine* 68, no. 12, 2009, 2215-22.

<sup>27</sup> See for example the report of the third Caldicott Review: NATIONAL DATA GUARDIAN FOR HEALTH AND CARE: Review of Data Security, Consent and Opt-Outs 2016.

medical records available for research on an opt-out basis. The scheme's aim was to tap into the vast potential resource that the comprehensive and structured patient health records kept within the NHS represented, in order to improve health care delivery and promote "world-class health services research" in England. The data was to be collected and curated by a public governmental body, the Health and Social Care Information Centre, which would then control access to the data and assess applications for its use according to a structured system of governance, including review for sensitivity of information. There was no direct cost to patients and everyone would be included, unless they chose to opt out.

Across the Atlantic, the Institute for Systems Biology's "100k Wellness Project"<sup>28</sup> aims to use 'big data' approaches to analyse genomic, metabolomic and physiological data from 100,000 individuals to understand factors contributing to health and identify early diagnostic markers for disease. Participants are to be recruited via its partner company Arivale, which offers direct-to-consumer personalised health advice and "wellness coaching" on a fee-for-service-basis. Arivale collects the data of interest from each of its customers in order to identify "actionable possibilities" that can "improve wellness or avoid disease". This data is then used, under the terms of Arivale's service, to contribute to the project dataset, with findings set to form the basis for further health predictions. Customers pay \$3499 for a 12-month 'membership'.

On the face of things, one might think a free-to-participate, inclusive, state-run health data service aiming to create a public research resource would seem like a more desirable proposition overall than a privately-owned, for-profit company harvesting data from only those who can afford to pay and feeding benefits back primarily to those subscribers. Yet care.data attracted such widespread concern and criticism from UK publics and other stakeholders that ahead of the planned roll-out in 2014, it was first suspended for six months and then quietly shelved in 2016. Arivale, on the other hand, seems still to be comfortably operational.

The contrast between these two examples highlights something of a crossroads in the development of health data research and how it might be realized in future. Is the social contract with respect to health data already failing? When it comes to the future of big health data, will the law of the free-market jungle govern, or the social contract of the well-regulated state? This is a serious ethical concern: the way in which big health data is operationalized has a direct bearing on our ability to realise the social benefits of health research, as well as how those benefits will be made available and to whom. Characterising participation as a consumer good rather than a public good has implications for justice in terms of who is included and who will benefit. If big health

---

<sup>28</sup> [http://www.nxtbook.com/nxtbooks/gen/clinical\\_omics\\_issue3/#/22](http://www.nxtbook.com/nxtbooks/gen/clinical_omics_issue3/#/22).

data operates principally on a 'pay-to-participate' basis, who will be the participants, and who will be excluded? This will have an impact on the relevance and applicability of findings to the wider population, and therefore on who is most able to benefit from health data research, with consequent effects for global health justice.

In order to address these issues, we need to understand what is behind these phenomena. Analysis of the events around *care.data* suggests that multiple failures contributed to the demise of the scheme: defects in trust, both actual and apparent; doubt about the extent to which it would serve the public interest; inadequate communication; and the unaddressed tension it created in the relationship between patients and primary care providers<sup>29</sup>.

The social contract paradigm implies that a corollary of research participation should be that the benefits of science flow back to the public. One of the main threats to its stability, therefore, is the perception that the system is not in fact operating for public benefit but to serve private interests. Particular worries may attach to the use of data by commercial for-profit companies<sup>30</sup>, though the motivations of scientists whose agendas are perceived to be self-interested or insufficiently transparent may also be seen as suspect<sup>31</sup>. In the case of *care.data*, the media also contributed to fomenting public concern, with one article framing the proposal as the NHS "selling patient data for commercial use"<sup>32</sup>.

Commercialisation, however, is not the only factor in play: how should we understand the pushback against *care.data* by contrast with the relative success of direct-to-consumer schemes such as Arivale? Obviously there are significant differences between attitudes towards health care and the health system between the UK and US; notably, health care is characterized much more as a consumer good in the US, versus a public good that is part of the state's responsibility in the UK. Perhaps, however, it is not only commercial use as such, but the intrusion of commercial interests into the citizen-state relationship that disrupts the social contract. The aims of *care.data*

---

<sup>29</sup> TAYLOR, MARK: "Information Governance as a Force for Good? Lessons to Be Learnt from Care.Data," SCRIPTed 11, no. 1, 2014, 1-8; CARTER, PAM, LAURIE, GRAEME T, and DIXON-WOODS, MARY: "The Social Licence for Research: Why Care.Data Ran into Trouble," Journal of Medical Ethics 41, no. 5, 2015, 404-9.

<sup>30</sup> HADDOW, GILL et al.: "Tackling Community Concerns About Commercialisation and Genetic Research: A Modest Interdisciplinary Proposal," Social Science and Medicine 64, no. 2, 2007, 272-82; KETTIS-LINDBLAD, A. et al.: "Genetic Research and Donation of Tissue Samples to Biobanks. What Do Potential Sample Donors in the Swedish General Public Think?," European Journal of Public Health 16, no. 4, 2006, 433-40.

<sup>31</sup> Evidence of this is seen particularly in the field of regenerative medicine, where scientists who oppose DTC sales of unproven treatments often receive vociferous criticism from patients accusing them of protecting their interests in their own research.

<sup>32</sup> TAYLOR, MARK: "Information Governance as a Force for Good? Lessons to Be Learnt from Care.Data".



expressed the benefits of health data research explicitly in terms of an economic agenda, rather than health care; might this have been out of keeping with the 'social license' required to legitimize such activity<sup>33</sup>?

How then can we attempt to reconcile the roles of consumer and participant when it comes to big health data? Clearly, further research is needed into the complex views, experiences and relationships this area entails. As a preliminary hypothesis, however, we might suggest one problem is that of ownership, in the sense of not just commercial interests in data but control. Whom is science by, and whom is it for? If publics feel excluded from participation in the conventional institutions of science, perhaps because of a perceived lack of control or an absence of desired opportunities for participation, they may turn to other outlets that supply such opportunities or offer more scope for the exercise of agency. This 'counter-hegemony' supported by consumer demand is having an impact on other areas of science and innovation<sup>34</sup>; the same potential exists with respect to big health data.

In the case of data science, lack of opportunity may apply especially to research where additional data is collected from selected participants, rather than re-use of existing data. Experience with genomic research projects, for example, suggests that there will be interest in participating from people who are not part of the target population for inclusion; for such would-be participants, the direct-to-consumer genetics industry may prove an appealing alternative.

Consumer contracts also offer a different mode of engagement to that available in the patient role. It is notable that Hood's idea of precision medicine characterizes participants as "health consumers"<sup>35</sup> rather than patients. Direct-to-consumer health data services are often presented as enhancing autonomy and fulfilling people's right to seek information about their health<sup>36</sup>: Arivale's customers are offered "personalized data, cutting-edge science and tailored

---

<sup>33</sup> CARTER, PAM, LAURIE, GRAEME T, and DIXON-WOODS, MARY: "The Social Licence for Research: Why Care.Data Ran into Trouble".

<sup>34</sup> For example stem cell science and regenerative medicine, see SALTER, BRIAN, ZHOU, Y., and DATTA, S.: "Hegemony in the Marketplace of Biomedical Innovation: Consumer Demand and Stem Cell Science," *Social Science and Medicine* 131, 2015, 156-63.

<sup>35</sup> HOOD, LEROY and FLORES, MAURICIO: "A Personal View on Systems Medicine and the Emergence of Proactive P4 Medicine: Predictive, Preventive, Personalized and Participatory".

<sup>36</sup> SCOTT, CHRISTOPHER T and DEFRANCESCO, L.: "Selling Long Life," *Nature Biotechnology* 33, no. 1, 2015, 31-40.

coaching” and invited to “unlock your data”<sup>37</sup>. Such initiatives tend also to be framed in terms of “citizen science”, using the language of ‘empowerment’<sup>38</sup>.

One question that we must confront, then, is why and how citizens might feel excluded from and disempowered with respect to public science, such that these messages find purchase in the private sector? If we are to shore up the social contract and ensure that big health data fulfills its potential in terms of public health, we need to pay more attention to the fourth ‘P’ of Hood’s “4Ps”<sup>39</sup>: *participatory* medicine.

What is required for meaningful participation in the context of big health data? Clearly, it is more than simply having one’s data included; other expectations about the personal and public benefits of research and who controls the research agenda are also important. The rhetoric of ‘citizen science’ has been invoked by both private and public health data initiatives to promote participation, either by increasing the desirability of the product or by appealing to a sense of civic duty. The liberal application of this term, however, encompasses a wide variety of “complex and multifarious proposed relationships between science, public goods, societal good, and public participation”<sup>40</sup>. Woolley and colleagues urge a deeper enquiry into how the concept of ‘citizen’ is deployed in these various instances and the relationships, rights and responsibilities it implies in each case; likewise, to develop a normative understanding of participation in data science will require us to explore the role of the ‘scientific citizen’ with respect to the social contract and what it should entail<sup>41</sup>.

## 4. Beyond health in big data

The focus of bioethics with respect to big data has thus far been mostly on health research: the uses of health records, population genomics, biobanking, patient-driven data mining. This is understandable: here perhaps lie the most obvious benefits and dangers. An ethics of big data, however, requires us to think beyond health to the wider possible applications of data science.

---

<sup>37</sup> <https://www.arivale.com/>.

<sup>38</sup> SCOTT, CHRISTOPHER T and DEFRANCESCO, L.: "Selling Long Life".

<sup>39</sup> HOOD, LEROY and FLORES, MAURICIO: "A Personal View on Systems Medicine and the Emergence of Proactive P4 Medicine: Predictive, Preventive, Personalized and Participatory".

<sup>40</sup> ———: WOOLLEY, J PATRICK et al.: "Citizen Science or Scientific Citizenship? Disentangling the Uses of Public Engagement Rhetoric in National Research Initiatives," *BMC Medical Ethics* 17, no. 1, 2016, 33.

<sup>41</sup> *Ibid.*

In this regard, we must overcome not only health-exceptionalism but research-exceptionalism about the uses of 'our' data. Whilst we are worrying about how our health care systems might be using our patient information to conduct research, other agents are acquiring and using all sorts of information about us, quite possibly for purposes that are less to our benefit than the sort of big health data research that is often talked about. These agents are often driven by commercial interests and possibly also political ones: for example, it seems that US immigration services have recently begun to request information on social media accounts to assist in determining who should be granted entry<sup>42</sup>.

The first lesson we should draw from this is that big data has ethical implications far beyond health: indeed, the conceptual basis of big data is to link health with various other spheres of existence.

Second, we need to question why we should immediately be inclined to treat research uses of data as suspect, more than other uses. In a now-notorious incident, Facebook conducted an experiment that involved selectively manipulating the contents of some users' news feeds to measure the effect of "emotional contagion", that is, the extent to which our moods are affected by those of others in our virtual proximity<sup>43</sup>. When this experimentation was revealed, public outcry swiftly followed: users were incensed at having been made the subject of research without their knowledge. The research even provoked an "editorial expression of concern" from PNAS, who noted that the study was "not fully consistent with the principles of informed consent and allowing participants to opt out"<sup>44</sup>. But is consent really at the heart of the ethical concerns over this research?

The Facebook experiment highlights the need for ethical approaches that are fit to deal with the new challenges of big data and social media<sup>45</sup>. To be clear, it is not that we shouldn't be

---

<sup>42</sup> NIXON, RON: "Visitors to the Us May Be Asked for Social Media Information," *New York Times*, 28 June 2016. <https://nyti.ms/2jPrKIP>, accessed

<sup>43</sup>COVIELLO, LORENZO et al.: "Detecting Emotional Contagion in Massive Social Networks," *PloS One* 9, no. 3, 2014, e90315; KRAMER, ADAM D, GUILLORY, J. E., and HANCOCK, J. T.: "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks," *Proceedings of the National Academy of Sciences of the United States of America* 111, no. 24, 2014, 8788-90.

<sup>44</sup> "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks".

<sup>45</sup> SCHROEDER, RALPH: "Big Data and the Brave New World of Social Media Research," *Big Data & Society* 1, no. 2, 2014, 2053951714563194; KAHN, JEFFREY P., VAYENA, E., and MASTROIANNI, A. C.: "Opinion: Learning as We Go: Lessons from the Publication of Facebook's Social-Computing Research," *Proceedings of the National Academy of Sciences of the United States of America* 111, no. 38, 2014, 13677-9; KLEINSMAN, JOHN and BUCKLEY, SUE: "Facebook Study: A Little

concerned about Facebook's actions in this case: any attempt to manipulate our moods potentially deprives us of agency. Our mistake, however, lies in thinking that our agency is restored as long as and only when we are given the opportunity to refuse or consent to this manipulation in the research context. Manipulating content on the basis of algorithms is already something that Facebook and other sites do, with tenuous implicit consent only in the form of users' agreement to the terms of service. If we accept that this is common practice, why then does it suddenly become wrong to study the effects of doing it? Is the objection simply based on our loss of agency in that we are being controlled unawares? Or is it also that the deprivation of agency is occurring deliberately and to facilitate someone else's gain – in this case, the researchers?

Either way, these concerns are not unique to the research context. Indeed, there is an argument to be made that the application of these techniques outside of research may have far more drastic and concerning effects on global society. The year 2016 saw a dramatic right-wing turn of world political events, including the election of Donald Trump as US President and the 'Brexit' referendum. Now, suggestions have begun to emerge that both results may have been subject to a concerted campaign of voter manipulation, masterminded by a data analytics company using "micro-targeting" to deliver individualized political content<sup>46</sup> – what Jonathan Albright describes as a "micro-propaganda machine" producing a "fake news ecosystem"<sup>47</sup>. The company in question, Cambridge Analytica, has been said to use a combination of data analytics and psychological profiling based on people's social media and online activity "to precisely target individuals, to follow them around the web, and to send them highly personalised political messages"<sup>48</sup>.

---

Bit Unethical but Worth It?," *Journal of Bioethical Inquiry* 12, no. 2, 2015, 179-82. GRIMMELMANN, JAMES: "The Law and Ethics of Experiments on Social Media Users," *Colorado Technology Law Journal* 13, 2015, 219-72.

<sup>46</sup> CADWALLADR, CAROLE: "Google, Democracy and the Truth About Internet Search," *The Observer*, 4 December 2016. <https://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook>, accessed 1 June 2017; ———: "Robert Mercer: The Big Data Billionaire Waging War on Mainstream Media," *The Observer*, 26 February 2017. <https://www.theguardian.com/politics/2017/feb/26/robert-mercero-breitbart-war-on-media-steve-bannon-donald-trump-nigel-farage>, accessed 1 June 2017; ———: "The Great British Brexit Robbery: How Our Democracy Was Hijacked," *The Observer*, 1 May 2017. <https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy>, accessed 1 June 2017.

<sup>47</sup> ALBRIGHT, JONATHAN. "The #Election2016 Micro-Propaganda Machine." 2016. <https://medium.com/@d1gi/the-election2016-micro-propaganda-machine-383449cc1fba>, accessed 1 June 2017; ANDERSON, BERIT. "The Rise of the Weaponised Ai Propaganda Machine.." 2017. <https://medium.com/join-scout/the-rise-of-the-weaponized-ai-propaganda-machine-86dac61668b>, accessed 1 June 2017.

<sup>48</sup> CADWALLADR, CAROLE: "Google, Democracy and the Truth About Internet Search."

The extent to which this sort of activity has occurred deliberately and influenced public views or political processes may never be entirely known. The story is not yet over; some of the investigative journalism pieces in which these suggestions were made are now the subject of legal action by Cambridge Analytica and associated persons. Nevertheless, it is clear that the combination of social media, psychometrics and big data has potentially vast power to affect our world in ways of which we are still not fully aware.

Understanding what we do with data, how algorithms affect the way we intersect with the data stream, and the consequences this may have is vital if we are to take control of, and responsibility for, those effects. Research is necessary to develop this understanding. Looked at another way, given that our social media feeds are and will continue to be controlled by algorithms the effects of which we may not fully understand, it would be irresponsible not to do research on this. The Facebook Terms of Service, to which each user must agree in order to access the service, include consent to use of data for “internal operations” such as service improvement and research. These activities may potentially produce very important findings about the effects of social media. Objecting to them being made ‘research’ in the academic sense with the results made public seems nonsensical; do we want companies to keep useful knowledge to themselves? In short, the data is there; we cannot prevent others from making use of it. Especially given this, it would be irresponsible not to make use of it for beneficial purposes.

Secondly, focusing on individual consent as the lodestone of ethical permissibility in big data research is misdirected, primarily because it fails to capture and protect the range of interests we have in relation to research. Especially when it comes to big data research and its broader effects on society, we are all invested in the results, whether or not we choose individually to participate. If, on the basis of research that does not include your data (because you have not consented), decisions are made that limit your participation in society or unjustly constrain your possible ways of being in the world, your agency is nonetheless impaired. Being able to refuse participation on an individual basis, to say “not with my data”, is not an adequate remedy for this.

Examples such as those discussed above illustrate that the most serious ethical concerns in relation to big data go beyond the level of individual control and data privacy. As these cases show, even data that we do not consider private can be used in ways that we may not understand or approve of, and that may have harmful effects. In shaping an ethical approach to big data research, therefore, we should focus less on individual capacity to act as gatekeepers of our own information and more on the collective stewardship of a joint resource. We should spend less time trying to say what research can’t be done with our data and more time worrying about what else is being done with it, as well as what research can and should be done in order to achieve social benefits.

## 4.1 Towards an ethics of data

What is different about big data, about its 'bigness', is partly the new approach to knowledge-making that it may be seen to represent. Some have characterized this as 'data-driven' versus 'hypothesis-driven', noting that the world of big data is one in which "data-driven decisions are poised to augment or overrule human judgment"<sup>49</sup>. What ethical challenges, then, does the new epistemology of big data present?

First and most obvious, we need to be alert to unexpected, undesirable or unjust consequences of using big data. If health assessments are based on big data, for example, are the results of those assessments fair and what we expect of a just health care system? Although human judgment may not factor directly into the data analytics process, there is still scope for its application in evaluating the results and whether those results are ethical.

We need also to be aware of how processes and infrastructures around big data invisibly constrain and shape knowledge-making. The conceit of the 'big data' research approach is that it looks at everything, or at least a sufficiently large and unbiased subset of everything to produce an inherently objective and complete view. In actual fact, however, as Leonelli has shown in relation to biological data, certain sorts of data may be more tractable to becoming part of Big Data, meaning that such datasets give the "illusion of completeness" rather than actually being complete.<sup>50</sup> While this is not necessarily the case for all big data research, contrasting the stated epistemological approach of big biological data with the reality shows that we ought at least to be conscious of this dissonance and alert to its potential effects.

Next, we need to return to the idea of agency and with it responsibility. The advent of big data approaches to research moves us from a world of scientific practice in which "objects have agency" to one in which data and data handling processes also have agency. Who, though, takes responsibility for the exercise of that agency? There is a moral lacuna potentially implicit in the epistemology of big data: the idea that 'meaning makes itself' may falsely absolve us from responsibility for creating that meaning.

Aaron Levenstein is credited with an aphorism likening statistics to bikinis: "What they reveal is suggestive, but what they conceal is vital"<sup>51</sup>. In the same way, the supposed objectivity of

---

<sup>49</sup> FULLER, MICHAEL: "Big Data: New Science, New Challenges, New Dialogical Opportunities".

<sup>50</sup> LEONELLI, SABINE: "What Difference Does Quantity Make? On the Epistemology of Big Data in Biology", 7.

<sup>51</sup> RATCLIFFE, SUSAN, ed. *Oxford Essential Quotations*, 3rd ed. (Published online DOI: 10.1093/acref/9780191804144.001.0001: Oxford University Press, 2015).

big data can conceal crucial things behind the “opaque and automated”<sup>52</sup> process of algorithmic decision-making. Who can be held accountable for decisions, and what scope will there be for critical review of those decisions, when they are the product of vast machine ‘intelligence’ operating by processes beyond the capability of the human brain? This will be doubly problematic when it comes to machine-directed algorithmic evolution, that is, when the process of refining and improving the algorithms themselves is also handled by computers.

We also need to be critical about what kind of data we use to “show us how the world is”. On 3 February 2017, for example, those monitoring the hashtags trending on Twitter would have been justified in believing that a horrific attack had just taken place in the town of Bowling Green; the hashtag #bowlinggreenmassacre resulted, however, from an ‘alternative fact’ cited by Kelly-Ann Conway about an event that never happened.

Another example of how algorithmic big data interpretation can inadvertently lead to inaccurate representations is demonstrated by examining what machines learn from the data they are given. To illustrate this, we may turn once again to Google, whose autocomplete function uses algorithmic analysis together with string combinations and browsing patterns gathered from previous searches to try to predict what users are looking for. This seemingly useful function can even shortcut the need for the actual search: if one is looking for the correct spelling of a foreign word, or the wording of a common saying, the autocomplete suggestion often helpfully provides the answer directly. What people search for and click on, though, may represent a different world to the one we know and expect. Further, because we are inclined to place some stock in the autocomplete function as telling us something about how the world is, what our machines feed back to us has the power to shape our perceptions. Reports of autocomplete suggestions such as “Are women... evil?” and “Are Muslims... bad?”<sup>53</sup> paint a disturbing picture of the world as Google sees it, and shows it to us<sup>54</sup>.

More than ever, therefore, we need a social epistemology of big data to reveal how the “facts” emerging from big data are shaped by underlying social structures and practices. In relation to research integrity specifically, we need an account of scientific responsibility that is adequate to deal with big data science and the diffuse distribution of responsibility that it generates. Finally, to identify and grapple with the new challenges posed by machine learning and

---

<sup>52</sup> ZARSKY, TAL: “The Trouble with Algorithmic Decisions,” *Science, Technology, & Human Values* 41, no. 1, 2016, 118-32.

<sup>53</sup> CADWALLADR, CAROLE: “Google, Democracy and the Truth About Internet Search.”

<sup>54</sup> Though, for accuracy, it must be noted that the present author attempted to replicate this and received the comparatively innocent suggestion “Are women’s... razors taxed?” from Google UK on 15 February 2017.

algorithmic intelligence, we need a robust exploration of the issues associated with the ethics and governance of algorithms: from the data to which they are applied and its inherent biases, to the embedded values that they may re-inscribe, to their potential social impacts<sup>55</sup>.

## 4.2 Living in the world of Tlön

A final, more philosophical issue that the age of big data prompts us to ponder is how we should live in the age of disembodied data. Hayles refers to the process of “how data lost its body” as “becoming posthuman”<sup>56</sup>; what does it mean to be human when our existence in this world is as much virtual as real, depends as much on data as physical embodiment?

The world of quasi-virtual, data-driven reality is a world in which facts become increasingly changeable, ‘fake news’ and alternative knowledge can acquire apparent truth value through the workings of the data machine and the virtual reality of the world described by disembodied data can seep through into physical, real-world existence.

In the imagined world of Borges’ Tlön<sup>57</sup>, “esse est percipi”: objects are brought into existence by belief, or “become effaced and lose their details when they are forgotten.” This applies to places as well as objects. Borges writes: “A classic example is the doorway which survived so long as it was visited by a beggar and disappeared at his death. At times some birds, a horse, have saved the ruins of an amphitheater...” In our world, the increasing datafication of everything presages a time in which “the world will be Tlön”. When data is reality, reality becomes mutable. Our perceptions, shaped by the data we receive from the online world, can create reality and reify the virtual. Conversely, when we cease to perceive something, when the data available to us do not reflect its existence, in a way it ceases to exist.

Take Google Maps, for example, as a form of virtual location data. It may seem nonsensical to claim that the existence or non-existence of a place in Google Maps has any bearing on its real-world physical permanence: Google Maps is only a decade or so old, and most of us have first-hand experience of physical places existing well before this. But will they *continue* to exist, now

---

<sup>55</sup> ZIEWITZ, MALTE: "Governing Algorithms," *Science, Technology, & Human Values* 41, no. 1, 2016, 3-16; NEYLAND, DANIEL: "Bearing Account-Able Witness to the Ethical Algorithmic System," *ibid.*, 50-76; TUFEKCI, ZEYNEP: "Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency," *Colorado Technology Law Journal* 13, 2015, 203-18.

<sup>56</sup> HAYLES, N KATHERINE: *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics* Chicago: University of Chicago Press, 1999.

<sup>57</sup> BORGES, JORGE LUIS: "Tlön, Uqbar, Orbis Tertius," in *Labyrinths: Selected Stories & Other Writings*, ed. Donald A Yates and James E Irby (Available online <https://libraryofbabel.info/borges/labyrinths.pdf>: New Directions, 1964).



that the world has become GoogleMap? As we rely increasingly on information from the virtual world to allow us to navigate the physical one, does a place to which we cannot navigate still exist? Certainly, businesses that are “unMappable” will soon cease to be viable, at least in areas where letting our scrolling-fingers do the walking supersedes foot traffic.

It is not a great leap to think from how the data world influences the existence of places to our existence as persons. Our ability to participate in society is increasingly dependent on our virtual existence as data subjects. Consider the difficulty of opening a bank account, renting an apartment or securing employment without documentation, official proof of identity and a social security number: without our data, there is a sense in which we do not exist.

The world of *Tlön* is not only fictional but meta-fictional: within Borges’ story, *Tlön* is the world of the mythology of the itself-invented country of Uqbar, that begins to manifest in the real world. It is, appropriately as a metaphor for the data age, an invented fiction that becomes real via the world’s collective enthusiasm for “the minute and vast evidence of an orderly plan”. Confronted with a new and different way of seeing the world, “almost immediately, reality yielded on more than one account. The truth is that it longed to yield.” To what invented reality might we be yielding in our enthusiasm for big data? We may think that facts describe or represent the world; Borges’s tale reveals to us that the representations we make, or allow to be made, create the world. While the disciples of *Tlön* are “enchanted by... a rigor of chess masters, not of angels”; “a labyrinth devised by men”, we may be in danger of succumbing to the opposite fallacy, believing that big data and computers will reveal to us the divine laws of an orderly reality that humans alone cannot grasp.

Borges’ narrator asks, “Who are the inventors of *Tlön*?” In the world of big data, we may well ask: who makes those representations? Within the story of *Tlön*, it is a society of secret elites, the *Orbis Tertius*, toiling for generations, who reshape reality. In our new world of big data, who will be the “*tlonistas*”, and what power will they wield? The example of Cambridge Analytica and its influence on world politics serves as a cautionary tale in this regard.

To conclude, then: the bigness of big data in one sense is that it has the potential to dwarf humanness, to subsume our individuality; it is bigger than any one of us. Where, then, does each one of us fit, in the world of big data? How does big data reposition us as individual human persons?

The power of big data lies in seeing the collective picture. We often talk about the importance of the bigger picture, and not being able to see the wood for the trees — but we must not lose sight of the trees for the wood; that is to say, we must not fail to see the individuality of persons amongst the big data. Grouping our data allows us to make powerful inferences, but

lumping us together as an inseparable mass may fail to respect our value as persons. Big data calls for a new ethics of information that must both recognize the power of the collective and respect the value of the individual.

## References

- ◆ ACADEMY OF MEDICAL SCIENCES. *Personal Data for Public Good: Using Health Information in Medical Research.*, Academy of Medical Sciences, London 2006.  
*Realising the Potential of Stratified Medicine 2013.*
- ◆ ALBRIGHT, J. "The #Election2016 Micro-Propaganda Machine." 2016.  
<https://medium.com/@d1gi/the-election2016-micro-propaganda-machine-383449cc1fba>, accessed 1 June 2017.
- ◆ ANDERSON, B. "The Rise of the Weaponised Ai Propaganda Machine." 2017.  
<https://medium.com/join-scout/the-rise-of-the-weaponized-ai-propaganda-machine-86dac61668b>, accessed 1 June 2017.
- ◆ BORGES, J.L. "Tlön, Uqbar, Orbis Tertius." Translated by James E Irby. In *Labyrinths: Selected Stories & Other Writings*, edited by Donald A Yates and James E Irby. Available online <https://libraryofbabel.info/borges/labyrinths.pdf>: New Directions, 1964.
- ◆ BUTLER, D. "When Google Got Flu Wrong." *Nature* 494, no. 7436 (Feb 14 2013): 155-6.
- ◆ CADWALLADR, C. "Google, Democracy and the Truth About Internet Search." *The Observer*, 4 December 2016. <https://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook>, accessed 1 June 2017.
- ◆ ———. "The Great British Brexit Robbery: How Our Democracy Was Hijacked." *The Observer*, 1 May 2017 2017.  
<https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy>, accessed 1 June 2017.
- ◆ ———. "Robert Mercer: The Big Data Billionaire Waging War on Mainstream Media." *The Observer*, 26 February 2017 2017. <https://www.theguardian.com/politics/2017/feb/26/robert-mercer-breitbart-war-on-media-steve-bannon-donald-trump-nigel-farage>, accessed 1 June 2017.
- ◆ CAPLAN, A.L. "Is There a Duty to Serve as a Subject in Biomedical Research?" [In eng]. *IRB: A Review of Human Subjects Research* 6, no. 5 (Sep-Oct 1984): 1-5.

- ◆ CARTER, P., LAURIE, G.T., and DIXON-WOODS, M. "The Social Licence for Research: Why Care.Data Ran into Trouble." *Journal of Medical Ethics* 41, no. 5 (May 2015): 404-9.
- ◆ CHAN, S., and HARRIS, J. "Free Riders and Pious Sons--Why Science Research Remains Obligatory." *Bioethics* 23, no. 3 (Mar 2009): 161-71.
- ◆ CHAN, S., HARRIS, J., and SULSTON, J. "Science and the Social Contract: On the Purposes, Uses and Abuses of Science." In *Common Knowledge: The Challenge of Transdisciplinarity*, edited by Jerome Billotte, Moira Cockell, F Waldvogel and F Darbellay, 45-59. Lausanne: EPFL Press, 2010.
- ◆ COVIELLO, L, SOHN, Y., KRAMER, A. D., MARLOW, C., FRANCESCHETTI, M., CHRISTAKIS, N. A., and FOWLER, J. H. "Detecting Emotional Contagion in Massive Social Networks." *PloS One* 9, no. 3 (2014): e90315.
- ◆ CUKIER, K., and MAYER-SCHOENBERGER, V. "The Rise of Big Data: How It's Changing the Way We Think About the World." *Foreign Affairs* 92, no. 3 (2013): 28-40.
- ◆ DESMOND-HELLMANN, S. "Toward Precision Medicine: A New Social Contract?". *Science Translational Medicine* 4, no. 129 (Apr 11 2012): 129ed3.
- ◆ DIXON-WOODS, M., ASHCROFT, R. E., JACKSON, C. J., TOBIN, M. D., KIVITS, J., BURTON, P. R., and SAMANI, N. J. "Beyond "Misunderstanding": Written Information and Decisions About Taking Part in a Genetic Epidemiology Study." *Social Science and Medicine* 65, no. 11 (Dec 2007): 2212-22.
- ◆ DIXON-WOODS, M., and TARRANT, C. "Why Do People Cooperate with Medical Research? Findings from Three Studies." *Social Science and Medicine* 68, no. 12 (Jun 2009): 2215-22.
- ◆ DIXON-WOODS, M., WILSON, D., JACKSON, C., CAVERS, D., and PRITCHARD-JONES, K. "Human Tissue and 'the Public': The Case of Childhood Cancer Tumour Banking." *BioSocieties* 3, no. 1 (2008): 57-80.
- ◆ DOWELL, S.F., BLAZES, D., and DESMOND-HELLMANN, S. "Four Steps to Precision Public Health." *Nature* 540 (2016): 189-91.
- ◆ FULLER, M. "Big Data: New Science, New Challenges, New Dialogical Opportunities." *Zygon* 50, no. 3 (2015): 569-82.
- ◆ GRIMMELMANN, J. "The Law and Ethics of Experiments on Social Media Users." *Colorado Technology Law Journal* 13 (2015): 219-72.

- ◆ GROVES, P, KAYYALI, B, KNOTT, D, and KUIKEN, V. *The 'Big Data' Revolution in Healthcare: Accelerating Value and Innovation*, Centre for US Health System Reform Business Technology Office 2013.
- ◆ HADDOW, G, LAURIE, G., CUNNINGHAM-BURLEY, S., and HUNTER, K. G. "Tackling Community Concerns About Commercialisation and Genetic Research: A Modest Interdisciplinary Proposal." *Social Science and Medicine* 64, no. 2 (Jan 2007): 272-82.
- ◆ HARRIS, J. "Scientific Research Is a Moral Duty." [In eng]. *Journal of Medical Ethics* 31, no. 4 (Apr 2005): 242-8.
- ◆ HAYLES, N.K. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago: University of Chicago Press, 1999.
- ◆ HENRICH, J., HEINE, STEPHEN J., and NORENZAYAN, A. "Most People Are Not Weird." *Nature* 466, no. 7302 (Jul 01 2010): 29.
- ◆ ———. "The Weirdest People in the World?". *Behavioral and Brain Sciences* 33, no. 2-3 (Jun 2010): 61-83; discussion 83-135.
- ◆ HOOD, L., and FLORES, M. "A Personal View on Systems Medicine and the Emergence of Proactive P4 Medicine: Predictive, Preventive, Personalized and Participatory." *New Biotechnology* 29, no. 6 (Sep 15 2012): 613-24.
- ◆ HORNE, R., BELL, J. I., MONTGOMERY, J. R., RAVN, M. O., and TOOKE, J. E. "A New Social Contract for Medical Innovation." *Lancet* 385, no. 9974 (Mar 28 2015): 1153-4.
- ◆ KAHN, J.P., VAYENA, E., and MASTROIANNI, A. C. "Opinion: Learning as We Go: Lessons from the Publication of Facebook's Social-Computing Research." *Proceedings of the National Academy of Sciences of the United States of America* 111, no. 38 (Sep 23 2014): 13677-9.
- ◆ KETTIS-LINDBLAD, A., RING, L., VIBERTH, E., and HANSSON, M. G. "Genetic Research and Donation of Tissue Samples to Biobanks. What Do Potential Sample Donors in the Swedish General Public Think?". *European Journal of Public Health* 16, no. 4 (Aug 2006): 433-40.
- ◆ KLEINSMAN, J., and BUCKLEY, S. "Facebook Study: A Little Bit Unethical but Worth It?". *Journal of Bioethical Inquiry* 12, no. 2 (2015): 179-82.
- ◆ KRAMER, A.D., GUILLORY, J. E., and HANCOCK, J. T. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Sciences of the United States of America* 111, no. 24 (Jun 17 2014): 8788-90.

- ◆ LEONELLI, S. "What Difference Does Quantity Make? On the Epistemology of Big Data in Biology." *Big Data & Society* 1, no. 1 (2014): 1-11.
- ◆ LLÀCER, M.R., CASADO, M., and BUISAN, L. *Document on Bioethics and Big Data: Exploitation and Commercialisation of User Data in Public Health Care*, Observatori de Bioètica i Dret, Barcelona 2015.
- ◆ MAYER-SCHÖNBERGER, V., and CUKIER, K. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray, 2013.
- ◆ MESLIN, E. M., and CHO, M. K. "Research Ethics in the Era of Personalized Medicine: Updating Science's Contract with Society." *Public Health Genomics* 13, no. 6 (2010): 378-84.
- ◆ NATIONAL DATA GUARDIAN FOR HEALTH AND CARE. *Review of Data Security, Consent and Opt-Outs* 2016.
- ◆ NEYLAND, D. "Bearing Account-Able Witness to the Ethical Algorithmic System." *Science, Technology, & Human Values* 41, no. 1 (2016): 50-76.
- ◆ NIXON, R. "Visitors to the Us May Be Asked for Social Media Information." *New York Times*, 28 June 2016. <https://nyti.ms/2jPrKIP>, accessed
- ◆ NUFFIELD COUNCIL ON BIOETHICS. *The Collection, Linking and Use of Data in Biomedical Research and Health Care: Ethical Issues* 2015.
- ◆ PRECISION MEDICINE INITIATIVE (PMI) WORKING GROUP REPORT TO THE ADVISORY COMMITTEE TO THE DIRECTOR, NIH,. *The Precision Medicine Initiative Cohort Program – Building a Research Foundation for 21st Century Medicine*. September 17, 2015.
- ◆ RATCLIFFE, S., ed. *Oxford Essential Quotations*. 3rd ed. Published online DOI: 10.1093/acref/9780191804144.001.0001: Oxford University Press, 2015.
- ◆ RHODES, R. "Rethinking Research Ethics." *American Journal of Bioethics* 5, no. 1 (Winter 2005): 7-28.
- ◆ SALTER, B., ZHOU, Y., and DATTA, S. "Hegemony in the Marketplace of Biomedical Innovation: Consumer Demand and Stem Cell Science." *Social Science and Medicine* 131 (Apr 2015): 156-63.
- ◆ SCHROEDER, R. "Big Data and the Brave New World of Social Media Research." *Big Data & Society* 1, no. 2 (2014): 2053951714563194.
- ◆ SCOTT, C.T, and DEFRANCESCO, L. "Selling Long Life." *Nature Biotechnology* 33, no. 1 (Jan 2015): 31-40.

- ◆ TALBOT, D. "Cell-Phone Data Might Help Predict Ebola's Spread." *MIT's Technology Review*, August 22 2014. <https://www.technologyreview.com/s/530296/cell-phone-data-might-help-predict-ebolas-spread/>, accessed 1 June 2017.
- ◆ TAYLOR, M. "Information Governance as a Force for Good? Lessons to Be Learnt from Care.Data." *SCRIPTed 11*, no. 1 (2014): 1-8.
- ◆ TUFEKCI, Z. "Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency." *Colorado Technology Law Journal 13* (2015): 203-18.
- ◆ VAYENA, E., BROWNSWORD, R., EDWARDS, S. J., GRESHAKE, B., KAHN, J. P., LADHER, N., MONTGOMERY, J., et al. "Research Led by Participants: A New Social Contract for a New Kind of Research." *Journal of Medical Ethics 42*, no. 4 (Apr 2016): 216-9.
- ◆ WALL, M. "Ebola: Can Big Data Analytics Help Contain Its Spread?" *BBC News*, 15 October 2014. <http://www.bbc.co.uk/news/business-29617831>, accessed 1 June 2017.
- ◆ WOOLLEY, J.P., MCGOWAN, M. L., TEARE, H. J., COATHUP, V., FISHMAN, J. R., SETTERSTEN, R. A., JR., STERCKX, S., KAYE, J., and JUENGST, E. T. "Citizen Science or Scientific Citizenship? Disentangling the Uses of Public Engagement Rhetoric in National Research Initiatives." *BMC Medical Ethics 17*, no. 1 (Jun 04 2016): 33.
- ◆ ZARSKY, T. "The Trouble with Algorithmic Decisions." *Science, Technology, & Human Values 41*, no. 1 (2016): 118-32.
- ◆ ZIEWITZ, M. "Governing Algorithms." *Science, Technology, & Human Values 41*, no. 1 (2016): 3-16.

**Fecha de recepción: 5 de junio de 2017**

**Fecha de aceptación: 30 de junio de 2017**