HSS VI.3 (2017)

DOI: 10.1515/hssr -2017-0022

Big Data for Biomedical Research and Personalised Medicine: an Epistemological and Ethical Cross-Analysis

Mathieu Guillermin*, Thierry Magnin**
Université Catholique de Lyon, France

Abstract
Big data techniques, data-driven science and their technological applications raise many serious ethical questions, notably about privacy protection. In this paper, we highlight an entanglement between epistemology and ethics of big data. Discussing the mobilisation of big data in the fields of biomedical research and health care, we show how an overestimation of big data epistemic power – of their objectivity or rationality understood through the lens of neutrality – can become ethically threatening. Highlighting the irreducible non-neutrality at play in big data tools, we insist upon the ethical importance of a critical epistemological approach in which big data are understood as possibly valuable only when coupled with human intelligence and evaluative rationality.

Keywords
big data, data-driven science, biomedical research, personalised medicine, precision medicine, epistemology, ethics

1. Introduction

During the past decades, the emergence of "big data" algorithms and technologies have deeply reconfigured the manner in which we relate to each other and to our environments, notably through social, business or

* UMRS-449, Post-doctoral student, the Catholic University of Lyon; mguillermin@univ-catholyon.fr
** UMRS-449, Chancellor of the Catholic University of Lyon; tmagnin@univ-catholyon.fr

governance processes. One of the key aspects of big data techniques, besides the large amount of data they allow to process, is "a capacity to search, aggregate, and cross-reference" large and differently structured datasets (Boyd and Crawford, 2012: 663). Even more characteristic is the dissociation that these new tools trigger between data collection and storage on one side, and data treatments and their purposes on the other side. Different actors such as citizens, consumers, private corporations or public institutions form information ecosystems that they feed through massive data collection which are stored without unified structuration or precisely determined treatments and purposes (Caseau, 2014). Purposes in constant mutation can then be addressed through dedicated algorithms (sometimes cultivated, in an agile manner, through machine learning) that reconstruct specifically structured data, enabling contextually adapted value creation.

Big data have numerous applications in domains ranging from e-commerce and market intelligence to governance, security, and public safety. Such digital tools are also highly beneficial in "data-intensive" scientific research, where enormous amounts of data have to be manipulated. In particular, big data are fruitfully mobilised in the biomedical domain (Leonelli, 2012). Digital epidemiology constitutes a particularly striking application (Salathé et al., 2012), as exemplified by the famous though mitigated Google Flu Trend experience aiming at the prediction of flu activity through web search patterns (Lazer et al., 2014). Moreover, big data processing of patient-specific information (from electronic health records, aggregated clinical trials, biomolecular sequencing data,…) permits the design of clinical decision support systems and opens the road for personalised or precision medicine (Duffy, 2016).

However, the rapid growth of big data techniques and applications raises serious ethical concerns, notably about privacy protection or massive surveillance, but also regarding algorithm transparency (especially when machine learning is involved), data ownership, digital divides, trust between data subjects and data processors, or intellectual property and access rights (Mittelstadt et al., 2016). Although laws and regulations progressively adapt to such challenges, as with the US "Customer Privacy Bill of Rights" (2012) or with the new EU regulation (2016/279) and directive (2016/280), careful "ethical foresight" is indispensable to avoid dogmatic postures and whiplash effects (Mittelstadt and Floridi, 2016 b: 1).

In this paper, we focus upon a specific type of ethical concerns that the deployment of big data techniques in biomedical research and health care may generate. The hypothesis we explore is that misinterpreting biomedical big data at the epistemological level can undermine the contributions such disruptive techniques could make, by rendering them ethically problematic. More precisely, we defend the idea that uncritical overestimation of the epistemic power of big data – for instance by conceiving data-driven science as a radical answer to the limitations of hypothesis-driven scientific research that irreducibly embeds human subjectivity – may lead to problematic reductionism and foster excessive medicalisation (Conrad and Schneider, 1980; Vogt et al., 2016). In addition, it may undermine the valuable contribution that big data tools could make to the practice of medicine, leading to automation instead of genuine personalisation (in which the traditional practice of medicine and the role of physicians in the therapeutic partnership with their patients would be preserved and enriched).

To support this claim, we draw on our involvement in the ethical dimensions of the French consortium Metagenopolis and of the EU research project MetaCardis[1] that mobilise big data techniques to elucidate the roles of gut microbiota in human health and disease. In section 2, we detail the manner uncritical overestimation of big data epistemic power can render ethically problematic the outcomes of such biomedical research. In section 3, we criticisethis type of overestimation and insist upon the indispensability of evaluative rationality. This leads us to highlight, in section 4, the crucial role of ethics and, more broadly, of human intelligence in big data-assisted biomedical research and health care.

2. Overestimating big data's epistemic power: an ethical issue

The reflection we propose in this paper upon the ethical significance of the epistemology of big data takes its source in our contribution to the ethical layer of the ongoing MetaCardis research project. As its name suggest, MetaCardis studies Cardio-Metabolic Disorders or Disease (CMD ranging from obesity and diabetes to atherothrombotic diseases and heart failure). CMD constitute an extremely acute medical, social and economic challenge that will become more and more problematic as population ages.

They are at the centre of a complex web of influencing factors among which one can mention individual's genome as well as environmental parameters (such as pollution or in utero environment) and lifestyle specificities (e.g. physical or dietary habits). However, the mechanisms structuring this web remain to be elucidated. Drawing upon recently discovered influences of dietary habits upon the composition and diversity of gut microbiota and its microbiome or metagenome, MetaCardis explores the hypothesis that alterations of the latter can trigger a modification of a patient's metabolic functioning susceptible to cause CMD's symptoms (Dao et al., 2016). This approach is extremely original in that it goes beyond the commonly accepted consideration of the sole host's genome, to take into account the way host's genome expression is altered by other components like the gut metagenome.

To probe the role of gut microbiota and its metagenome, MetaCardis deploys a big data approach through a multilevel framework rooted in systems biology and systems medicine. In virtue of the complexity of phenomena associated with CMD, the approach can be nothing but *big* data. Big data techniques are expected to permit the extraction of insightful correlations, of "small patterns" (Floridi, 2012), from a spectrum of relevant data that is extremely broad and diversified. MetaCardis involves different types of subjects (CMD patients at different stages, patients with different predisposing factors and control groups) from which various sorts of information are collected. Clinical, anthropometrical, and biochemical data (like weight, glucose tolerance status, or adiposity) are integrated with information obtained through questionnaires about medical history and lifestyle (dietary habits and physical activity, but also socio-economic status or educational level, and psychological stress or perceived quality of life,…). These various data are complemented with the collection of biosamples allowing high-throughput molecular sequencing to establish "omics" profiles of subjects involved in the study. Thereby, a *metagenomic* characterisation of the gut microbiota composition and diversity can be coupled with *metabolomic* analysis of molecules involved in metabolic processes, and with *transcriptomic* monitoring of host's genome expression. Moreover, harmonizing and integrating such complex datasets, as well as detecting relevant correlations with respect to CMD constitute open methodological challenges. The project thus comes with the objective of building its own data processing software to integrate and visualise data, and to compute models. Statistical and machine learning approaches are developed for patient stratification or correlation detection.

This ambitious and powerful framework of investigation is expected to improve the understanding of CMD. A first goal is to settle refined understanding for these disorders or diseases by identifying metabolomic CMD markers, in addition to already known inflammatory patterns and other biomedical data. A second target is the elucidation of metagenomic CMD markers (gut microbiota signatures) that are correlated with these metabolomic markers. These evidenced connections should permit a modelling of patient's metabolic functioning that accounts for the role of the gut microbiota. In turn, such a modelling would permit to simulate the impact of gut microbiota alterations upon a patient's metabolic functioning. Such progresses may even lead to redefine CMD stages by identifying (new) pathways of CMD progression in the global metabolic functioning. This would open ground-breaking perspectives in translational research and health care. Diagnostics and prognostic biomarkers could be defined in "omics" terms (disease-associated microbiomes, metabolomics-derived markers, transcriptomic signatures), complementing the more traditional pathophysiological targets. In addition, exploiting the various types of information available in the study's datasets, new therapeutic approaches could be proposed that would aim at restoring or achieving specific gut microbiota configurations through prescriptions upon lifestyle or environmental factors, notably upon dietary habits (Cotillard et al., 2013). In consequence, biomedical states of patients could be categorized in terms of omics-based CMD criteria, pathological categories being associated with therapeutic prescriptions (especially upon lifestyle or environmental factors). Coupled with high-throughput omics-sequencing of patients, the road would be opened for personalized or precision medicine, with individually designed patient care.

To maximize the chances for such very appealing perspectives to deliver their full potential, MetaCardis comes with an embedded dimension of ethical questioning and foresight we contribute to. Among the ethical issues explored, one concerns the manner in which big data techniques and their applications are understood at the epistemological level. In fact, philosophical positions that uncritically overestimate big data neutrality and objectivity may induce problematic ethical consequences. This is, for instance, the case with data fundamentalist approaches that conceive big data-based processes of knowledge production as the highest forms of objectivity and rationality, to the extent that they are allegedly free from human subjectivity or biased interpretations. According to data fundamentalism, "massive data sets and predictive analytics always reflect objective truth," and "unless you have really large datasets then you're not getting close to objectivity and truth" (Crawford, 2013a; Crawford, 2013b). Some data fundamentalists even argue that "with enough data, the numbers speak for themselves" (Anderson, 2008). This data fundamentalist narrative promises us the emergence of an *exclusively* "data-driven science," which would produce absolutely objective knowledge by rendering unnecessary the involvement of human intelligence, with its irreducible load of interpretation and subjectivity (Mittelstadt and Floridi, 2016a: 462-464).

Such epistemological positions that are extremely "optimistic" about the epistemic power of big data raise several issues, especially when they are mobilised to inform decision-making and action. Generally speaking, data fundamentalism may prevent one from suitably considering the possibility of mistake, the validity of big data results being assumed rather than genuinely demonstrated (Wigan and Clarke, 2013: 47-48). Big data results are understood as definitive and infallible, which undermines the possibility to seek for progress and improvement. Focusing on the field of biomedical research and health care, the ethical investigation we contribute to, in the framework of MetaCardis, highlights more specific concerns that, if left unaccounted for, could undermine the value of the project's outcomes.

First, the idea that big data techniques deploy the highest form of rationality, neutrally providing objective truth, may promote the belief that big data analytics result in absolute and exhaustive representations of reality superseding any other description. On this ground, this risk of "ontic occlusion" – "the process by which emphasizing particular aspects of a phenomenon in a discourse necessarily occludes or 'downplays' other aspects" (Mittelstadt and Floridi, 2016a: 470) – is not negligible. Although a methodological form of reductionism (focusing on specific aspects of a problem for pragmatic reasons) is often desirable (Magnin and Revol, 2015: 59), the overestimation of big data's epistemic power would open the door to a more radical form of reductionism, denying the very existence of any other aspects than those captured by big data results. Relevant aspects that are not reflected in mobilised data could thereby be overlooked. On the one hand, some aspects may be neglected because their signature is masked in the noise or simply because some information is missing. The inclusion of additional data would then be susceptible to resolve the issue. On the other hand, there might be relevant dimensions that are not amenable to digital data. These would be discarded by principle. As a matter of illustration, the ethical foresight within MetaCardis indicates that ontic occlusion could occur about patients' subjective experiences, although this dimension may prove to be ethically crucial. In fact, authors such as Michel Henry highlight vulnerability and affectability of lively beings in their radical immanence as an ontological feature of life that is at the root of the possibility for human beings to be genuine persons (Magnin, 2015). Occulting such dimensions would deprive ethical thinking from any ground to resist reifying technological interventions that threaten patients' autonomy (Magnin and Revol, 2015: 59-62).

Insofar as one of the main reasons big data techniques are disruptive is their capability to gather many different and heterogenous types of data, the risk of reductionism triggered by an overestimation of big data epistemic power may be somehow counterbalanced. But, even when big data results capture all relevant dimensions of the problem under scrutiny (notwithstanding the problem of dimensions that are not amenable to digital data), data fundamentalism could still generate other issues. In addition to the question of reductionism, the overestimation of the epistemic power of big data techniques may lead to a second threat: it could encourage excessive "medicalization" (Conrad and Schneider, 1980). Medicalization designates "the expansion of medical jurisdiction." It occurs when medical vocabulary and conceptions are adopted to treat a new problem that was not considered as medical beforehand. With medicalization, problems that could also be considered, for instance, as social or psychological (such as alcoholism or hyperactivity) are approached through medical lenses and receives "a medical form of treatment (e.g. prescribing tranquilizers for an unhappy family life)" (ibid, 75-76). Biomedical big data, understood as the most objective form of biomedical investigation, are likely to radicalize such a process. The ethical investigation embedded in MetaCardis identifies over-medicalisation as a relevant topic. Correlating biological and medical phenomena with an extremely extended range of non-directly medical facts (such as patients' socio-economic status and educational level, their psychological stress, or the way they perceived the quality of their life) is a core strength of big data techniques mobilised in the project. It opens the possibility to coin therapeutic prescriptions in terms of these non-medical dimensions. But deprived from due ethical foresight, such valuable tools could lead to extreme forms of medicalization, with each person's whole life defined "in biomedical, technoscientific terms as quantifiable and controllable." Although it might not be the unique ground for it, data fundamentalism would surely foster and radicalize this type of over-medicalisation by blind trusting biomedical big data results as always reflecting perfectly objective truths.

This discussion of the topic of medicalisation naturally leads us to a third question, the one of personalised medicine, that is deeply impacted by the epistemological attitude toward big data techniques. Genuinely personalised medicine is one of the main targets of MetaCardis. High-throughput *omics* characterizations allow to position individuals with respect to big data-derived stratifications in terms of relevant CMD biomarkers, and associated with types and stages of disease (for diagnosis), with risks of their occurrence (for prognosis), and with suitable therapeutic prescriptions. According to MetaCardis ethical foresight, the extremely promising perspectives offered by such tools to enrich doctor-patient therapeutic partnerships can nonetheless be undermined with epistemological positions like data fundamentalism and its overestimation of big data'sepistemic power. To the extent that big data results are considered as absolutely objective, infallible, and exhaustive, the intervention of a physician with its subjectivity and human interpretations could be judged superfluous or even undesirable. Only engineers would be indispensable, as supervisors of algorithms functioning. Patient care could thereby become largely automatised. The temptation of automatised medicine could even become irresistible in contexts characterised by large cuts in public expenditures. But such an automation would fall short from genuinely personalised medicine. On the one hand, it would lead to downgrade or merely discard the role of the physician and its human intelligence. Infallibility and exhaustiveness render pointless any additional human intervention aiming at handling potential mistakes or at articulating big data results with other sources of information. On the other hand, there would be no room remaining for the role of the physician as a *genuine person*, and of the importance of the therapeutic *human relation* he or she establishes with patients (Mittelstadt and Floridi, 2016a: 471). In turn, this would also constitute failure at acknowledging and taking into account the specificities of individual patients as *persons*. The human dimension of the therapeutic partnership, its components that happen in one's heart of hearts, would be negated[2]. In addition, automatised medicine could even degenerate in extreme forms of control with the implementation of connected devices ensuring real-time monitoring of the biomedical state of patient, of their behaviour and their adherence to prescribed treatments. Although new techniques to evaluate treatment adherence could be valuable when implemented within a doctor-patient fiduciary relationship, it is necessary to reject any automated framework that would severely threaten patients' autonomy by systematically privileging big-data derived representations of situations at stakes over more human points of view.

In sum, the work of ethical foresight undertaken within MetaCardis leads to consider as a crucial ethical topic the manner big data techniques are understood at the epistemological level. In particular, we should resist the uncritical admission of the claim that big data-driven investigation instantiates, in virtue of its neutrality, the highest form of rationality and objectivity.

3. Data-fundamentalism as an epistemological illusion: non-neutrality and evaluative rationality

The idea that neutrality is a core component of objectivity and rationality is not new. It is central in received understandings of scientific investigation. Rooted in empirical-formal approaches of scientific method, these received views share the belief in the existence of a universal scientific method permitting to infer knowledge claims from empirical evidence and tools of logic. This scientific method is considered as neutral insofar as only logic and empirical evidence matter for theory justification. Subjective and contextual specificities are irrelevant with respect to theory justification. However, this core belief of received views has been deeply criticised during the 20th century, notably with the Duhem-Quine thesis about confirmational holism, and with the post-positivist stream famously represented by the works of Kuhn or Feyerabend. For these authors, the scientific method is not neutral. Scientific investigations irreducibly mobilize non-neutral elements that can vary in function of historical, cultural or societal contexts. Consequently (methodological), incommensurability can occur when researchers rooted in different contexts do not share a common measure for theory justification (Oberheim and Hoyningen-Huene, 2013). Nonetheless, this post-positivist understanding of scientific method has itself been intensively criticised as leading to relativism and undermining science rationality (Baghramian, 2014). On this ground, the attraction that operates the idea of data fundamentalism is not a surprise. The belief that big data techniques, and the data-driven science they enable, could overcome the lack of neutrality of the usual scientific method revives the idea of rationality and objectivity qua neutrality.

But it seems that this narrative about the objectivity of big data techniques is closer to mythology than to sound epistemology (Boyd and Crawford, 2012: 663; Mittelstadt and Floridi, 2016a: 462). Big data processing is far from neutral. Hypotheses and commitments are indispensable at several levels. First of all, data themselves, regardless of their size, are not neutral. "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care" (Bowker 2005, quoted in: Boyd and Crawford, 2012: 663). For instance, one cannot avoid "design decisions that determine what will be measured" (Boyd and Crawford, 2012: 667). Data are influenced by tools used for their acquisition. They provide "views from certain vantage points, rather than an all-seeing, infallible God's eye view." Synthetically put, "data are created within a complex assemblage that actively shapes its constitution" (Kitchin, 2014: 4-5). In biomedical research, one could mention, for instance, data about a patient's blood pressure, which highly depends on contexts and techniques of measurement (data acquired through medical records with a physician, or through real-time monitoring wearable devices; data acquired as the patient rest, or just after a physical effort, …). In sum, any database is built in a specific non-neutral way (choices in categories, techniques of collection or measurement, methods for data sampling and curation, procedures for handling missing data,…) that depends on the context (Canali, 2016; Mittelstadt and Floridi, 2016a: 464-465). These non-neutral elements can prove crucial for the correct interpretation of embedded data. This non-neutrality of databases' construction is a first reason to resist any mythological overestimation of big data's epistemic power. In addition, this contextual aspect of any database implies that the process of aggregating diverse types of information sources, which is one of the strengths of big data techniques, is far from straightforward. It is not mere juxtaposition. Database harmonisation also requires specific decisions that depends upon final goals and on initial types of sources. In MetaCardis, such non-neutrality is acknowledged. Databases' construction and harmonisation are considered as open methodological questions that require cautious treatment for big data processing to be epistemically robust.

Second, no results can be established through the algorithmic processing of aggregated big datasets without non-neutral commitments. The mythology we criticise in this section claims that big data techniques are particularly efficient to detect small patterns in huge amounts of data, patterns that could not have been identified otherwise (Floridi, 2012). As a preliminary reminder, it is worth mentioning that such patterns are not yet theories, causal links or modelling bringing to the fore mechanisms and explanations. One can think of the well-known problem of under-determination of theory choice. Here it is under-determination with respect to correlations. In fact, one can imagine that the same set of evidenced correlations can be associated with different possible (causal) mechanisms when trying to model phenomena underlying these correlations. Additional judgments or commitments (for instance in favour of some epistemic values such as simplicity), as well as pre-admitted theoretical or causal knowledge, will thus be required to opt in favour of one of these competing mechanisms. Correlations alone cannot do the full job here (Canali, 2016). This being said, the core of the mythology about the epistemic power of big data is not yet debunked insofar as it often comes with "an empiricist mode of knowledge production" in which correlation is enough (Kitchin, 2014: 4-5). It is precisely, we are told, because insights can be gained directly from data without any theoretical hypothesis that big data techniques reach a higher form of objectivity and rationality. Big data do not need to elucidate (causal) mechanisms. Instead they provide purely objective "actionable insights" (Mittelstadt et al., 2016: 3).

Notwithstanding potential resisting issues with actionable insights[3], it is primary here to point out that even correlation settling through big data techniques is not neutral (Kitchin, 2014; Mittelstadt and Floridi, 2016a: 462-465). The inductive strategy deployed through such techniques to identify patterns "does not occur in a scientific vacuum." Data can be framed following many different systems of categories and can be treated by the mean of various data-mining techniques. Noise elimination together with datasets boundaries redefinition can also be performed in many ways. In addition, data processing and visualising tools are developed on the ground of specific pre-existing computational, mathematical and statistical approaches. For example, through suitable non-neutral decisions, multivariate methods can be deployed to decipher manageable patterns based on a reduced number of principal components or variables (Canali, 2016: 6). Therefore, big data cannot operate without "theoretically informed decisions" that rely on particular values, previous findings, pre-admitted theories and scientific approaches, or past experience (Kitchin, 2014: 5-6). In MetaCardis, data processing steps of this type (which demand non-neutral decision) are carefully managed. For instance, results of previous studies of CMD (MetaHit consortium, Nut Obese, Microbaria, Micro-Obes,…) are used to suggest potentially relevant biomarkers and gut microbiota species, as well as to establish the pre-stratification of patients. Interestingly, MetaCardis also relies upon machine learning to evidence relevant biomedical patterns. But machine learning is not understood as a way to restore pure neutrality. Generally speaking, machine learning implementation also requires non-neutral decisions, which demands cautious analysis. In some cases, decisions are required to define targets (trial-error or genetic algorithms). In other cases, algorithms are trained on previously generated data that imply their own share of non-neutrality. As Kate Crawford recently claimed[4], biases hosted in human-generated data easily propagate to trained algorithms: "We should always be suspicious when machine learning systems are described as free from bias if it's been trained on human-generated data," Crawford said. "Our biases are built into that training data."

In sum, big data techniques do not fundamentally modify the topic of neutrality of scientific research. Non-neutral decisions are indispensable, not only in traditional (hypothesis-driven) investigations as pointed by Kuhn and the post-positivists, but also in big data-enabled research as we just recall. In this respect, usual scientific research and data-driven science are on a par. Does this mean that (even big data-enabled) scientific research is irremediably plagued by subjectivity and that relativism or irrationalism cannot be escaped? The question is delicate, but the important point is that this radical conclusion follows only if one strictly equates non-neutrality with irrationality or lack of objectivity. And such a connection is far from straightforward. In his famous work *Reason, Truth and History,* Putnam proposes an interesting approach to this question (Putnam, 1981: chapters 6, 8, and 9). He holds that scientific inquiries (in particular those of hard sciences) are "central examples of rational thinking" (p. 135). Therefore, doubting the rationality of science is not the right answer to the irreducibility of non-neutrality. Instead, we should try to understand and to resist our tendency to systematically condemn non-neutrality. And for Putnam, we tend to worry about non-neutrality because it implies choices, decisions, or judgments whose validity is not trivial and requires *evaluation*. For instance, the scientific method necessitates epistemic principles, such as simplicity, that we believe we *have to* follow (Putnam, 1981: chapitre 6). Similarly, the validity of non-neutral decisions in big data needs evaluation. We must ask ourselves: are such decisions *good*? According to this analysis, non-neutrality puzzles us because we tend to discard the possibility of rationality about evaluative matters.

For Putnam, this tendency is largely rooted in the extremely widespread fact-value dichotomy (which can be traced back to Hume and the origin of empiricism). This dichotomy leads us to believe that only the establishment of facts can be fully rational and objective – insofar as only facts can be perceptively verified, though sometimes only indirectly. By contrast, judgments about values (which cannot be observed) – and by extension, all evaluative claims – are expelled from the realm of rationality and objectivity (Putnam, 2002: 102). In consequence, once the fact-value dichotomy is accepted, *rational evaluation* is impossible. Non-neutrality is then associated with lack of objectivity or of rationality, and (data-driven) science rationality is undermined (threat of relativism). But this reasoning can be avoided. Putnam proposes robust arguments against the dichotomy (Putnam, 2002). For him, nothing prevents, in principle, evaluative rationality. He even offers some insights about the face evaluative rational inquiries could take (Putnam, 2004). Without detailing too much, Putnam indicates first that evaluative rational inquiries can be guided, as any rational inquiries, by (epistemic) principles such as Kant's categorical imperative, Habermas' ethics of discourse or Dewey's fallibilism and democracy (Putnam, 2008: 385-387; Putnam, 2004: 10 and 25). In addition, one could even consider the possibility of empirical testing for evaluative discourses. In effect, once the fact-value dichotomy is abandoned, nothing forbids connecting evaluative claims to observational ones (Putnam, 2002: 14-15).

Whatever this may be, the important point Putnam makes is that evaluative rationality is possible. Consequently, non-neutrality in (data-driven) science can be acknowledged without opening the door to relativism or irrationalism. Rationality and objectivity are not systematically achieved by seeking absolute neutrality (which may well be an illusion). Rather, researchers can increase the rationality of their investigations by conducting (reflexive) inquiries upon their practices of investigation (Popa et al., 2015). Rationality and objectivity can be gained through the identification and rational evaluation (accompanied with, if required, modification) of non-neutral decisions these practices irremediably embed. As we elaborate in the next section, this conclusion means that it is misled to understand the elimination of human intelligence as an efficient tool to pursue rationality and objectivity. On the contrary, we have just highlighted the indispensability of evaluative judgments performed by researchers themselves, on the ground of their fallible, though rational, tools.

4. Biomedical big data, human intelligence, and ethics: toward a genuinely personalised medicine

As detailed in the previous section, the ethical foresight embedded in MetaCardis highlights the indispensability of human intelligence and evaluative rationality for fruitful big data-enabled biomedical research and (personalised) health care. This acknowledgement comes with two important valuable ethical consequences.

To begin with, it is not only epistemically, but also ethically crucial to correctly grasp the epistemological status of big data techniques. Broadly speaking, the acknowledgement of non-neutrality and fallibilism of big data tools, and of the results they produce, is indispensable to keep in mind that epistemic robustness is not automatic. The suitable algorithmic processing of a huge amount of data requires human intelligence and its readiness to criticise and improve established (big data) practices by deploying evaluative rationality. This general point is of great significance to defuse ethical threats delineated in section 2 (reductionism, overmedicalisation, and dehumanised medicine).

In fact, the irreducibility of non-neutrality (and the consecutive indispensability of human intelligence and evaluative rationality) means that big data-derived results deserve no absolute superiority. By principle, room is always open for other legitimate insights. Thereby, reductionism loses most of its power of attraction. When addressing a given question through data-driven research, it becomes clear that obtained results depend upon a set of non-neutral decisions. As such, they are not necessarily exhaustive. Other sets of non-neutral commitments may also be admissible and may lead to other complementary findings. Moreover, once epistemological absolutising of big data tools is rejected, room is also secured for the legitimate consideration of aspects that might not be amenable to big data treatment. This insight is reflected in the framework of MetaCardis through the clear understanding that results in terms of metagenomic, metabolomic or transcriptomic markers (potentially coupled to clinical data and to information about patients' life) do not necessary exhaust what should be said about CMD. Such results can be recognised as highly valuable without negating the interest of other discourses, for instance about subjective factors.

The same line of thought permits circumventing the ethical threat of over-medicalisation. As we have just indicated, big data results involving biomedical concepts are neither infallible nor necessarily exhaustive. Therefore, correlations can be established between biomedical states of affair and elements of other dimensions of patients' life (such as their educational level or the way they perceive the quality of their life), without claiming absolute supremacy for the former over the latter. Some correlations may be misleading. Even when they are robust, superiority cannot be granted to medical characterisation by principle. Human judgment remains indispensable to determine, in function of effective problematic configurations, whether it is legitimate or not to consider through medical lenses given aspects of persons' life. Instantiated within

MetaCardis framework, this leads to value the possibility to understand dietary habits as a medical problem with respect to CMD, without discarding the consideration of these habits in non-medical perspectives. In this respect, the notion of functional food is interesting. Carefully designed food can be a powerful therapeutic tool to fight or prevent CMD. Still, this medical perspective may legitimately cohabitate with, for example, reflection upon the social and cultural status of food in sustainability issues, regarding which medical dimensions are not central (Vivero-Pol, 2017).

In addition, the acknowledgment of the irreducibility of non-neutrality and of the role of human intelligence (with its evaluative rationality) is a key element MetaCardis relies on to reach the genuine personalisation of medicine, instead of an ethically threatening automation and dehumanisation. Big data tools coupled with fast sources of individual information (as high-throughput "omics") do permit to position each patient in categories associated with diagnosis or prognosis markers, and with suitable treatments. However, these individualised conclusions are not understood as absolute and infallible truths. Moreover, such a categorisation of patients is not necessarily an exhaustive account of phenomena at stake. Therefore, it becomes crucial to recognise the irreducibility of the role of physicians (or health care professionals) with their human intelligence, and with the associated ability to rationally evaluate situations of individual patients, as well as to gauge contributions of big data techniques in consequence.

First, results from big data techniques being fallible, physicians are indispensable to assess their epistemic robustness and to avoid possibly harmful blind application. Second, the human intelligence of physicians is also necessary to articulate data-based conclusions with other insights about patients understood as complex and thick entities, as persons with their subjectivity. It is one of the roles of physicians to resist data reductionism by maintaining room open for other aspects that may escape big data-assisted analysis. Physicians are required to determine, case by case, which aspects deserve participating therapeutic practices and the manner these aspects should be combined. Finally, because attention and legitimacy is granted to subjective experiences of patients, one can preserve the idea that the affectability and emotional skills of physicians as human persons are major and irreducible components of therapeutic partnerships.

This dimension is notably crucial to address ethical issues related to patients' autonomy. As a matter of illustration, we have already mentioned the topic of treatment adherence. Big data tools (possibly coupled with real-time monitoring wearable devices) may permit to ensure that patients respect prescribed treatments. In an automated framework, such verifications could directly lead to measures of control and to sanctions. By contrast, once the role of physicians is acknowledged, monitoring tools of this type can be suitably employed to signal when something goes wrong in the therapeutic partnership (thereby providing the opportunity to deepen doctor-patient collaborations to overcome actual obstacles). In sum, preserving the role of human intelligence and rational evaluation in therapeutic partnerships is a key element to escape the threat of a dehumanising automation of medicine, which would exclusively focus on individuation along bio-physiological parameters. This human and evaluative layer is indispensable for big data-based individualisation to reach genuine personalisation.

Maintaining room explicitly open for human intelligence and evaluative rationality is thus a crucial ethical step. But this is only a first step, which leads to a second dimension: ethical issues also reside in the answers human intelligence brings in performing non-neutral decisions. Although some of them have exclusively epistemic implications, certain non-neutral decisions can also have an ethical impact. Accordingly, ethics should be part of rational evaluations guiding these non-neutral choices. In light of the ethical foresight embedded in MetaCardis, several loci requiring specific reasoning can be identified. Interestingly, these loci are not confined to the application of big data results or to the implementation of ready-made big data tools. Some may be found at the very beginning of data-driven research, embedded in the initial design of big data processing tools. In particular, ethical reasoning could be required about decisions for the initial framing of databases (e.g. concerning data selection, data curation, or choices of categories). For instance, although it may not be false by itself, structuring databases for correlation seeking in terms of ethnical or socio-economical categories may become ethically problematic (Mittelstadt and Floridi, 2016a: 467). Specific communities could capture potential benefits. Conversely, a negative outcome may impact only specific groups. Similarly, determining a level of relevance for noise elimination can have some ethical consequences. Important ethical concerns also arise when deciding whether individualised research results about individual participants in biomedical research such as MetaCardis should (or must) be communicated (or not) to participants, their physicians, or even to public health authorities (McGuire et al., 2012).

Another point at which ethical questioning appears necessary bears upon the manner the different tools available for (personalised) patient care are mobilised and combined. In this sense, we can mention the need to suitably regulate medicalisation. Finding the right balance between the importance granted to medical and biological accounts and the role of other dimensions may require subtle ethical reasoning. For example, we could imagine a CMD patient whose prescriptions in terms of dietary habits lead to psychological stress and dissatisfaction with life. Up to which point should such prescriptions prevail? In the same vein, one can wonder how individualised big data results should be weighted when taking extremely serious medical decisions. For instance, what is the place of individual results about big data-enabled prognosis (e.g. for a patient to develop a serious cardiometabolic condition) when deciding whether someone should receive an organ transplant? Although some of these ethical issues are not fundamentally original, the rise of big data-enabled biomedical research and health care may open new dimensions and needs to be anticipated. The case of personalised medicine is an adequate illustration. For it to become genuinely personalised, human intelligence should not be expelled from the picture. But this human intelligence should also be ethically informed to meet the specific challenges paving the road from the very beginning of data-driven biomedical research up to its implementation at the level of patient care.

5. Conclusion

In this paper, we drew upon the work of ethical foresight we contribute to within MetaCardis research project to evidence an entanglement between epistemology and the ethics of big data. We argued that the extremely promising perspectives opened by big data-enabled biomedical research and health care can be undermined by the ethical consequences of an uncritical overestimation of the epistemic power of big data. Data fundamentalism promotes such an overestimation by claiming that the algorithmic processing of massive amounts of data constitutes the highest form of rationality or objectivity, and produces absolute truths and purely neutral results. Data fundamentalism may have gained some traction as an answer to post-positivist criticisms of the neutrality of the traditional scientific method. However, we defended that it leads to an understanding of data-driven science – as a new form of radical empiricism that could be assimilated to naïve inductivism – that, not only fails to genuinely overcome the issue of science non-neutrality, but may also become ethically questionable.

The main lines of our argumentation have been the following:

- We reported on the ethical foresight we contribute to within the framework of the big data-enabled biomedical research project MetaCardis to show that data fundamentalism fosters serious ethical threats such as reductionism, over-medicalisation, or automated but dehumanised medicine.
- We then pointed that data fundamentalism is misled by evidencing the irreducible non-neutrality at play in many loci ranging from the design of big data tools to their application to patient care. This led us to question the soundness of epistemological views that assimilate rationality or objectivity with neutrality. Equating objectivity and rationality with pure neutrality is at best utopian, and possibly chimerical.
- Alternatively, we proposed an epistemological approach that acknowledges the irreducibility of non-neutrality and the indispensability of human intelligence. Human intelligence is necessary to perform irreducible non-neutral decisions and to evaluate them. In this approach, rationality and objectivity have to be conquered through reflexive investigations aiming at *rationally evaluating* research practices.
- On this ground, we argued that it is misleading and ethically problematic to value big data as a solution to eliminate the non-neutrality of human intelligence. On the contrary, our analysis showed that big data processing can valuably contribute to biomedical research and health care only when coupled to human intelligence and its ability to rationally perform (and evaluate) non-neutral decisions. Associated with human

intelligence and rational evaluation, data-driven science can then be properly understood as a powerful tool that may be legitimately and fruitfully combined with more traditional scientific practices to enrich the manner scientific knowledge is produced.

As we argued, the ethical consequences of this conclusion are twofold:

- First, recognising this irreducible role of rational evaluation is itself ethically crucial. It allows to better understand how to ensure the epistemic robustness of big data processing and its applications (this has indirect ethical importance). It also permits to defuse the general threats of reductionism, over-medicalisation, and dehumanised medicine. In effect, it shows that researchers and physicians with their human intelligence are indispensable.

- Second, it is not epistemically but also ethically crucial to carefully explicate and rationally evaluate (and, if required, to modify) the different non-neutral choices that are necessary to settle big data-enabled biomedical research, as well as to deploy its results for health care. This is so because some of these choices may embed direct ethical consequences. As such, their explication and rational evaluation is not only an epistemic matter. It should integrate ethical reasoning.

To conclude, the reflection proposed in this paper can be prolonged with a comment about the way ethics is understood. The last point we just recalled suggests that we should take distance from an overly restrictive vision of ethics as external criticism of research results, technologies' implementations, and the danger they may generate. We have just seen that ethical reasoning can also legitimately contribute to explanations and rational evaluations that enhance rationality or objectivity during the very design of research tools and practices, as well as during the development of technological applications. This suggests an original mode for ethical thinking: ethics in co-construction or embedded ethics. In our view, ethics in co-construction would be interdisciplinary by essence. It would gather scientists and experts of concerned fields with ethicists in collaborative processes aiming at enhancing rationality and objectivity of scientific research, technological developments, and consecutive applications within societies. When societal stakes are high, ethics in co-construction may even become transdisciplinary by integrating non-expert stakeholders (such as citizens, economic actors, members of patients' associations,…). Such ethics in co-construction may fruitfully contribute to the rationality of techno-scientific transformations of our societies, helping to drive them toward increased humanisation.

References

Anderson, C.(2008).“The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” Wired.06-23-2008, https://www.wired.com/2008/06/pb-theory/.

Baghramian, M. (2014).“Relativism About Science.”*The Routledge Companion to Philosophy of Science*.Eds. S. Psillos, and M. Curd.London – New York: Routledge. 268-279.

Boyd, D., and Crawford, K. (2012).“Critical Questions for Big Data.”*Information, Communication & Society*. 15.5: 662-679.

Canali, S. (2016). “Big Data, Epistemology and Causality: Knowledge in and Knowledge out in Exposomics.”*Big Data & Society*3.2:1-11.

Caseau, Y. (2014).“Big data : un changement de paradigme peut en cacher un Autre.”*Académie des Technologies de la Commission TIC Version 0.4*(Septembre 2014).

Conrad, P., and Schneider, J.W. (1980).“Looking at Levels of Medicalization: A Comment on Strong's Critique of the Thesis of Medical Imperialism.”*Social Science & Medicine. Part A: Medical Psychology & Medical Sociology* 14.1: 75-79.

Cotillard, A., et al. (2013).“Dietary Intervention Impact on Gut Microbial Gene Richness.”*Nature*. 500: 585-588.

Crawford, K. (2013a).“Big Data Becomes 2013's Scariest Fad.”*Future Tense*. 05-13-2013, http://www.abc.net.au/radionational/programs/futuretense/kate-crawford-big-data/4686244.

Crawford, K. (2013b). “The Hidden Biases in Big Data.”*Harvard Business Review*. 04-01-2013, https://hbr.org/2013/04/the-hidden-biases-in-big-data.

Dao, M.C., et al. (2016).“*Akkermansia Muciniphila* and Improved Metabolic Health During a Dietary Intervention in Obesity: Relationship with Gut Microbiome Richness and Ecology.” Gut 65.3 (2016): 426-436.

Duffy, D.J. “Problems, Challenges and Promises: Perspectives on Precision Medicine.” Briefings in Bioinformatics 17.3 (2016): 494-504.

Floridi, L. “Big Data and Their Epistemological Challenge.” Philosophy &

Technology 25.4 (2012): 435-437.

Kitchin, R. "Big Data, New Epistemologies and Paradigm Shifts." Big Data & Society1.1 (2014): 1-12.

Lazer, D., et al. "The Parable of Google Flu: Traps in Big Data Analysis."Science 343 (2014): 1203-1205.

Leonelli, S. "Introduction: Making Sense of Data-Driven Research in the Biological and Biomedical Sciences." Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences43.1 (2012): 1-3.

Magnin, T. "Vulnerability at the Heart of the Ethical Implications of New Biotechnologies." Human and Social Studies 4 (2015): 13-25.

Magnin, T., and Revol, F. "Life in Terms of Nano-Biotechnologies." Issues in Science and Theology: What Is Life?.Eds. D. Evers, et al.. Cham: Springer International Publishing, 2015. pp. 57-65.

McGuire, A.L., et al. "Perspectives on Human Microbiome Research Ethics."*Journal of Empirical Research on Human Research Ethics*, vol. 7, no. 3, 2012, pp. 1-14, doi:doi:10.1525/jer.2012.7.3.1.

Mittelstadt, B.D., et al. "The Ethics of Algorithms: Mapping the Debate."*Big Data & Society* 3.2 (2016): 1-21.

Mittelstadt, B.D., and Floridi, L. (2016a)."The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts."*The Ethics of Biomedical Big Data*.Eds. B.D. Mittelstadt, and L. Floridi.Cham: Springer International Publishing. 445-480.

Mittelstadt, B.D., and Floridi, L. (2016b). "Introduction."*The Ethics of Biomedical Big Data*. Eds. B.D. Mittelstadt, and L. Floridi. Cham: Springer International Publishing. 1-13.

Oberheim, E., and Hoyningen-Huene, P. (2013). "The Incommensurability of Scientific Theories."*The Stanford Encyclopedia of Philosophy*. Ed. E.N. Zalta, http://plato.stanford.edu/archives/spr2013/entries/incommensurability/.

Popa, F., et al. (2015)."A Pragmatist Approach to Transdisciplinarity in Sustainability Research: From Complex Systems Theory to Reflexive Science."*Futures* 65: 45-56.

Putnam, H. (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.

Putnam, H. (2002). *The Collapse of the Fact/Value Dichotomy: And Other Essays.* Cambridge – London: Harvard University Press.

Putnam, H. (2004). *Ethics without Ontology*. Cambridge – London: Harvard University Press.

Putnam, H. (2008). "Capabilities and Two Ethical Theories."*Journal of Human*

Development 9.3: 377-388.

Salathé, M., et al. "Digital Epidemiology."PLOS Computational Biology 8.7 (2012): 1-5.

Vivero-Pol, J. (2017). "Food as Commons or Commodity? Exploring the Links between Normative Valuations and Agency in Food Transition."*Sustainability* 9.3: 442.

Vogt, H., et al. (2016). "The New Holism: P4 Systems Medicine and the Medicalization of Health and Life Itself." *Medicine, Health Care and Philosophy* 19.2: 307-323.

Wigan, M.R., and Clarke, R. (2013). "Big Data's Big Unintended Consequences." *Computer* 46.6: 46-53.

[1] The Catholic University of Lyon (with its Group for Epistemology and Ethics of Science and Technology - GEEST) is co-founder of the Metagenopolis consortium (http://www.mgps.eu) whose members collaborate with different partner research projects such as MetaCardis (http://metacardis.net). Metagenopolis deals with metagenomic analyses of human gut microbiota, while MetaCardis aims to relate these analyses to cardio-metabolic disorders and disease.

[1] In light of this discussion, considering that big data tools may permit the personalisation of medicine appears rather inadequate. Traditional medicine already relies on a personalised doctor-patient partnerships. Therefore, a more valid approach would be to picture big data-enabled biomedical research as susceptible to enrich (or impoverish, in ethically problematic cases) the personalised dimension of medicine by permitting a more precisely individualised account of patients' biomedical state.

[2] On the one hand, the idea of *actionable insight*, understood as non-causal information, seems hardly sufficient in research fields expected to inform future decision-making and action, as it is the case with biomedical research (Canali, 2016; Mittelstadt et al., 2016: 5). On the other hand, one can wonder to which extent the notion of 'actionable insight' is genuinely non-causal, if it is really meant to guide action as the label suggests.

[3] In an interview given to*The Guardian* (the 13th of March 2017): *Artificial intelligence is ripe for abuse, tech researcher warns: 'a fascist's dream'*, https://www.theguardian.com/technology/2017/mar/13/artificial-intelligence-ai-abuses-fascism-donald-trump.

Biographical notes

Mathieu Guillermin holds a Ph.D. in Physics (interaction laser-matter) and a Ph.D. in Philosophy of Science (pragmatist epistemology for combining incommensurability and context-sensitivity with scientific realism). He has recently joined the Catholic University of Lyon to study philosophical and ethical topics associated with digital technologies.

Thierry Magnin is both a Professor in Physics and in Ethics, and currently, the Rector of the Catholic University of Lyon, France. He has a Ph.D. in Physics (Solid State Physics) and a Ph.D. in Theology (Moral philosophy as a ground for a new dialogue between science and religion). He has written about 200 papers in solid state physics and several books on the ethics of biotechnologies: *Les nouvelles biotechnologies en question* (Salvator 2013); *Penser l'humain au temps de l'homme augmenté* (Albin Michel 2017). He has received a prize from the Academy of Science in France, and is a member of the French Academy of Technology.